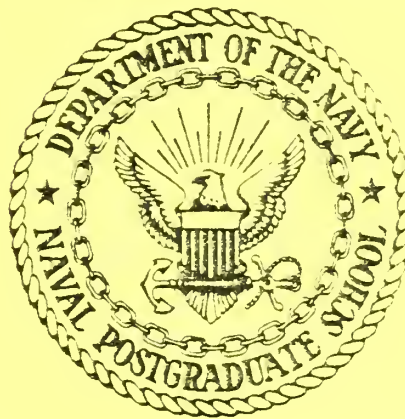


NPS55-78-036Pr

NAVAL POSTGRADUATE SCHOOL

Monterey, California



SMOOTHING 3-D DATA FOR TORPEDO PATHS

by

J. B. Tysver

May 1978

Approved for public release; distribution unlimited.

Prepared for: Research and Engineering Department
Naval Undersea Warfare Engineering Station
Keyport, Washington 98345.

FEDDOCS
D 208.14/2:
NPS-55-78-036PR

NAVAL POSTGRADUATE SCHOOL
MONTEREY, CALIFORNIA

Rear Admiral T. F. Dedman
Superintendent

J. R. Borsting
Provost

The work herein was supported in part by funds provided by
the Naval Undersea Warfare Engineering Station, Keyport, Washington.
Reproduction of all or part of this report is authorized.

This report was prepared by:

4 2 1

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE		READ INSTRUCTIONS BEFORE COMPLETING FORM
1. REPORT NUMBER NPS55-78-036Pr	2. GOVT ACCESSION NO.	3. RECIPIENT'S CATALOG NUMBER
4. TITLE (and Subtitle) Smoothing 3-D Data for Torpedo Paths	5. TYPE OF REPORT & PERIOD COVERED Technical	
	6. PERFORMING ORG. REPORT NUMBER	
7. AUTHOR(s) J. B. Tysver	8. CONTRACT OR GRANT NUMBER(s)	
9. PERFORMING ORGANIZATION NAME AND ADDRESS Naval Postgraduate School Monterey, Ca. 93940	10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS N0025378WR00002	
11. CONTROLLING OFFICE NAME AND ADDRESS Research and Engineering Department Naval Undersea Warfare Engineering Station Keyport, WA 98345	12. REPORT DATE May 1978	
	13. NUMBER OF PAGES 64	
14. MONITORING AGENCY NAME & ADDRESS (if different from Controlling Office)	15. SECURITY CLASS. of this report Unclassified	
	15a. DECLASSIFICATION DOWNGRADING SCHEDULE	
16. DISTRIBUTION STATEMENT (of this Report) Approved for public release; distribution unlimited.		
17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)		
18. SUPPLEMENTARY NOTES		
19. KEY WORDS (Continue on reverse side if necessary and identify by block number) Torpedo path Track smoothing Least squares		
20. ABSTRACT (Continue on reverse side if necessary and identify by block number) The general track smoothing program (MASM3DRJ) in use at NUWES uses linear, parabolic, and logarithmic functions to fit 3-D data files on torpedo paths by the method of least squares. Polynomial functions of the first (linear), second (parabolic), third, and fourth orders were fitted to data for a variety of path segments of a torpedo run at NUWES using the method of least squares. Results suggest expansion of the program to include higher order polynomials and fitting shorter path segments will provide substantial reduction in residual errors. CONTINUED OVER		

DD FORM 1 JAN 73 1473

EDITION OF 1 NOV 65 IS OBSOLETE
S/N 0102-014-5601

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

20. Abstract cont.

The method of sequential differences was tried on the data and can be incorporated in the smoothing program as a means of identifying outlier data points and of selecting the appropriate polynomial order for fitting the data.

SMOOTHING 3-D DATA FOR TORPEDO PATHS

I. THE GENERAL PROBLEM

A. Data

Data in the form of ordered quadruplets $(t_i, x_i, y_i, \text{ and } z_i)$ are available from 3-D files on torpedo and target paths. The times t_i are sufficiently accurate so that they can be assumed to be without errors. The spatial coordinates $x_i, y_i, \text{ and } z_i$, however, are not only subject to measurement errors, but also may contain erratic measurements or have measurements missing for some of the equally spaced time intervals.

B. Desired Output

Information to be extracted from this data can be obtained either as:

(1) smoothed information as a function of time (parametric form), or

(2) smoothed information at a particular sequence of times which can be specified.

A comparison of computational requirements of the two procedures will involve the length of intervals used in smoothing and the number of times in the sequence of times of interest. Both procedures involve the same smoothing techniques.

The information to be extracted from the 3-D data includes:

(1) smoothed position coordinates

(a) as functions of time (i.e., $x=f_x(t), y=f_y(t), z=f_z(t)$)

(b) at specified times t_i (i.e., $x(t_i), y(t_i), z(t_i)$).

(3) velocity component estimates

(a) as functions of time (i.e., $V_x(t)$, $V_y(t)$, $V_z(t)$)

(b) at specified times t_i (i.e., $V_x(t_i)$, $V_y(t_i)$, $V_z(t_i)$),

(4) relative torpedo and target geometry in vicinity of intercept.

C. Data Sample

The path of the torpedo involves maneuvers so that segments must be selected for applications of the smoothing technique. The lengths of the segments, and hence the number of possible data points, is open to selection. Curves to be used to fit the data will primarily be polynomials. Longer path segments will generally require higher order polynomials and be more difficult to fit with acceptably small residuals. On the other hand, short intervals contain fewer data points and can limit capability for reducing prediction errors—the trade-off must be resolved by considering potential paths, and measurement errors. Some indication will be presented in subsequent sections of this report where data for a specific torpedo path is analyzed. Initially, two sample sizes ($n=11$ and $n=21$) are considered.

One of the questionable features for small sample sizes is possible further reduction by deletion of data points which appears inconsistent with the remaining data.

II. DATA SMOOTHING

A. Methodology

The data smoothing considered in this report is limited to the method of least squares. Other methods such as Kalman filtering would be appropriate for real time data smoothing where interest is centered on the next data point following the data used in the smoothing, but the current status of the method is not appropriate for postexperimental application where times within the data sample are of interest.

The data smoothing techniques currently used at IVPS involve the least squares method with the following equations:

$$(1) \hat{x}(t) = a + bt \quad (\text{linear})$$

$$(2) \hat{x}(t) = a + bt + ct^2 \quad (\text{quadratic, parabolic})$$

$$(3) \hat{x}(t) = a + b \ln(t) \quad (\text{logarithmic}).$$

This report concentrates on the addition of higher order polynomials, in particular:

$$(4) \hat{x}(t) = a_0 + a_1 t + a_2 t^2 + a_3 t^3 \quad (\text{cubic})$$

$$(5) \hat{x}(t) = a_0 + a_1 t + a_2 t^2 + a_3 t^3 + a_4 t^4 \quad (\text{quartic}).$$

The linear least squares technique is described in Appendix A. The sum of squares of the residuals

$$D = \sum_1^N e_i^2 = \sum_1^n \left(x_i - \hat{x}(t_i) \right)^2$$

provides a basis for selection of the particular equation to be used in fitting a particular set of data. The statistic

$$S_e^2 = D/(n-k),$$

where n is the number of points in the sample and k is the number of parameters in the equation, provides an estimate of the variance of measurement errors.

B. Sequential Differences

A preliminary screening of sample data by successive differences can serve a dual purpose:

(1) indication of the order of the polynomial required to produce a reasonable fit, and

(2) indication of isolated wild data points (outliers).

The first through fourth successive differences are presented in Table 1 when the actual relationship of x to t is linear and in Table 2 when the relationship is quadratic. A perturbation d is introduced in x_3 .

There are several salient features of successive differences that should be noted:

(1) Ignore, for the moment, the perturbation in x_3 . In Table 1, the first differences (the Δ_{1i} 's) consist of the velocity term a_1 plus noise. If a_1 is large with respect to the noise (the n_i 's), these differences will all have the same sign. The second differences (the Δ_{2i} 's); however, involve only noise and their signs should be random. This change from consistent signs for the Δ_{1i} 's to random signs for the Δ_{2i} 's is an indication that a linear relationship of x to t is appropriate.

In passing, it should be noted that:

$$\overline{\Delta}_1 = \frac{1}{6} \sum_{i=1}^6 \Delta_{1i} = a_1 + (n_6 - n_0)/6,$$

$$\frac{\sigma^2}{\overline{\Delta}_1} = \frac{\frac{\sigma^2}{n_6}}{36} + \frac{\frac{\sigma^2}{n_0}}{36} = \sigma^2/18,$$

$$n_i \sim N(0, \sigma^2)$$

$$x_i = x(t_i) = a_0 + a_1 t_i + n_i$$

t_i	x_i	Δ_{1i} $x_i - x_{i-1}$	Δ_{2i} $\Delta_{1i} - \Delta_{1, i-1}$	Δ_{3i} $\Delta_{2i} - \Delta_{2, i-1}$	Δ_{4i} $\Delta_{3i} - \Delta_{3, i-1}$
0	$a_0 + n_0$				
1	$a_0 + a_1 + n_1$	$a_1 + N_{11}$	N_{21}		
2	$a_0 + 2a_1 + n_2$	$a_1 + N_{12}$	$N_{22} + d$	$N_{31} + d$	$N_{41} - 4d$
3	$a_0 + 3a_1 + n_3 + d$	$a_1 + N_{13} + d$	$N_{23} - 2d$	$N_{32} - 3d$	$N_{42} + 6d$
4	$a_0 + 4a_1 + n_4$	$a_1 + N_{14} - d$	$N_{24} + d$	$N_{33} + 3d$	$N_{43} - 4d$
5	$a_0 + 5a_1 + n_5$	$a_1 + N_{15}$	N_{25}	$N_{34} - d$	
6	$a_0 + 6a_1 + n_6$	$a_1 + N_{16}$			
<hr/>					
	i	1	2	3	4
N_{ji}		$n_i - n_{i-1}$	$n_i - 2n_{i-1} + n_{i-2}$	$n_i - 3n_{i-1} + 3n_{i-2} - n_{i-3}$	$n_i - 4n_{i-1} + 6n_{i-2} - 4n_{i-3} + n_{i-4}$
$2\sigma_N$		$(1+1)\sigma^2 = 2\sigma^2$	$(1+2^2+1)\sigma^2 = 6\sigma^2$	$(1+3^2+3^2+1)\sigma^2 = 20\sigma^2$	$(1+4^2+6^2+4^2+1)\sigma^2 = 70\sigma^2$
σ_N		1.4σ	2.5σ	4.5σ	8.4σ
$3\sigma_N$		4.2σ	7.5σ	13.5σ	25σ
$3\sigma_N$		17	30	54	100
$\text{Max } \frac{ k d }{K \ 3\sigma_N}$		$\frac{d}{17}$	$\frac{d}{15}$	$\frac{d}{18}$	$\frac{d}{17}$
Critical Value		$a_1 + 17$	15	18	17

$$P(-3\sigma_N \leq N \leq 3\sigma_N) \geq .99$$

$$\sigma \doteq 4$$

Table 2. Sequential Differences - Quadratic Case

t_i	x_i	Δ_{1i}	Δ_{2i}	Δ_{3i}	Δ_{4i}
	$x_i = x(t_i) = a_0 + a_1 t_i + a_2 t_i^2 + n_i$		$n_i = N(0, \sigma^2)$		
0	$a_0 + n_0$	$a_1 + a_2 + N_{11}$	$2a_2 + N_{21}$		
1	$a_0 + a_1 + a_2 + n_1$	$a_1 + 3a_2 + N_{12}$	$2a_2 + N_{22} + d$	$N_{31} + d$	$N_{41} - 4d$
2	$a_0 + 2a_1 + 4a_2 + n_2$	$a_1 + 5a_2 + N_{13} + d$	$2a_2 + N_{23} - 2d$	$N_{32} - 3d$	$N_{42} + 6d$
3	$a_0 + 3a_1 + 9a_2 + n_3 + d$	$a_1 + 7a_2 + N_{14} - d$	$2a_2 + N_{24} + d$	$N_{33} + 3d$	$N_{43} - 4d$
4	$a_0 + 4a_1 + 16a_2 + n_4$	$a_1 + 9a_2 + N_{15}$	$2a_2 + N_{25}$	$N_{34} - d$	
5	$a_0 + 5a_1 + 25a_2 + n_5$	$a_1 + 11a_2 + N_{16}$			
6	$a_0 + 6a_1 + 36a_2 + n_6$				
Critical Value			$2a_2 + 15$	18	17

and that $\bar{\Delta}_1$ is normally distributed, i.e.,

$$\bar{\Delta}_1 \sim N(a_1, \frac{\sigma^2}{18}).$$

It should also be noted that if a_1 , is not large with respect to σ , the signs of the Δ_{1i} 's can still have the sign of a_1 with the dominance of this sign depending upon the relative magnitudes of a_1 and σ .

Next, consider the quadratic case (Table 2). The Δ_{3i} 's having random signs and the Δ_{2i} 's are dominated by the sign of a_2 , and hence the quadratics are indicated as the appropriate polynomial. Note that the signs of the Δ_{1i} 's may also be the same for all i if a_1 and a_2 have the same sign. If a_1 and a_2 have opposite signs and a_1 is greater than a_2 then there can be a change in the sign of the Δ_{nj} 's where $a_1 + (i^2 - (i-1)^2) a_2$ changes sign. In the vicinity of this point the n_j 's can become significant and produce some random sign terms.

Higher order differences are required to deal with higher order polynomials. In general, random signs in $(k+1)$ st order differences and consistent signs in k^{th} order differences indicate selection of a $(k+1)$ st order polynomial to fit the data.

(2) The perturbation d was included to provide an examination of the effect of an isolated outlier on successive differences. For illustrative purposes, it will be assumed that a successive difference greater than three times the standard deviation of the noise in that difference will be considered as an indication that a perturbation exists. The value $\sigma = 4$ will also be used for illustrative purposes.

Now, note the entries in the lower part of Table 1. Unless a_1 is known (or estimated) a critical magnitude for the Δ_{1i} 's cannot be specified. For higher order differences the i^{th} difference of the j^{th} order (Δ_{ji}) has a normal distribution.

$$\Delta_{ji} \sim N(k_{ji}d, \sigma_{N_j}^2)$$

Where k_{ji} is the coefficient of d in Δ_{ji} . If $d = 0$ then:

$$\Delta_{ji} \sim N(0, \sigma_{N_j}^2).$$

The situation is an application of statistical hypothesis testing. If Δ_{ji} is larger than can be expected due to noise alone, then the presence of a perturbation (an outlier) is indicated. The critical magnitude using assumptions of $1.0-0.99 = 0.01$ as significance level and $\sigma = 4$ is presented in the last row of Table 1. Thus if $|\Delta_{2i}| > 17$, $|\Delta_{3i}| > 18$, or $|\Delta_{4i}| > 17$, for any i , then an outlier is indicated.

Note that the value $\sigma = 4$ was assumed for this illustration. If sequential differences are used for preliminary screening before least squares curve fitting is performed, the estimate S_e for σ will not be available. A value of σ may be assumed from prior information of measurement errors but for purposes of preliminary screening some value greater than 4 would permit elimination of data points with large perturbations.

It should be emphasized that the above discussion pertains to the simplest situations. For applications where there are missing data points, or where perturbations are not isolated, more guidance will be required. The assumption that the noise components (the n_i 's) are independent and have the same variance, also warrants reservations in applications of the models.

III. APPLICATION

A. Sample Data

A specific test in which a torpedo was launched against a submarine at the Naval Undersea Warfare Engineering facilities will be used for illustration. The 3-D data includes equally spaced times from 814 to 1000—very few data points are missing. Figure 1 shows the torpedo path with every fifth point. Segments of this torpedo path are selected for application of the methodology presented in Section II. The presentation is restricted to the x and y coordinates.

B. Data Sample I

The initial 21 points (814-834) appear to lie in a straight line in Figure 1 and were selected as the first data sample. This data is presented in Figure 2 and Table 3.

(1) Successive differences:

The first and second order successive differences are also presented in Table 3. For the x component, all the first differences are negative and the second differences appear random (except possibly for the tail of the sample where a sequence of four pluses occur including one value ($\Delta_{2,17}=17.2$) which is large enough so that it might indicate an outlier). The alternating signs, (-, +, - or +, -, +) are not present so an isolated outlier does not appear likely.

For the y component, all the first order successive differences are positive and the second order differences appear somewhat random. Again, $\Delta_{2,17} = -13.2$ indicates that something has occurred in the vicinity of t_{18} . Higher order differences were not explored for this sample.

(2) Least squares smoothing:

Both linear and quadratic functions were fitted using the least squares method outlined in Appendix A. The results are presented below:

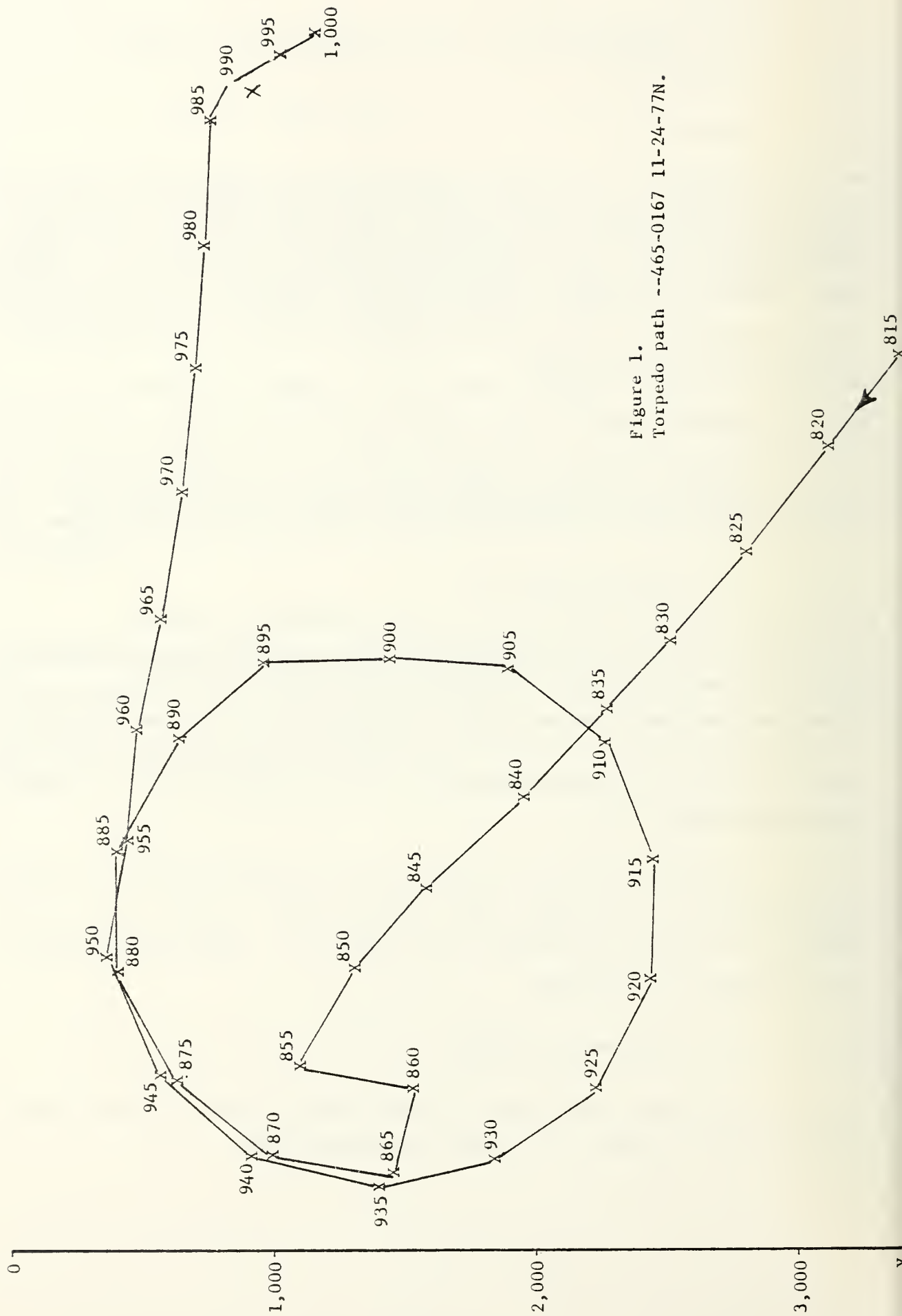


Figure 1.
Torpedo path --465-0167 11-24-77N.

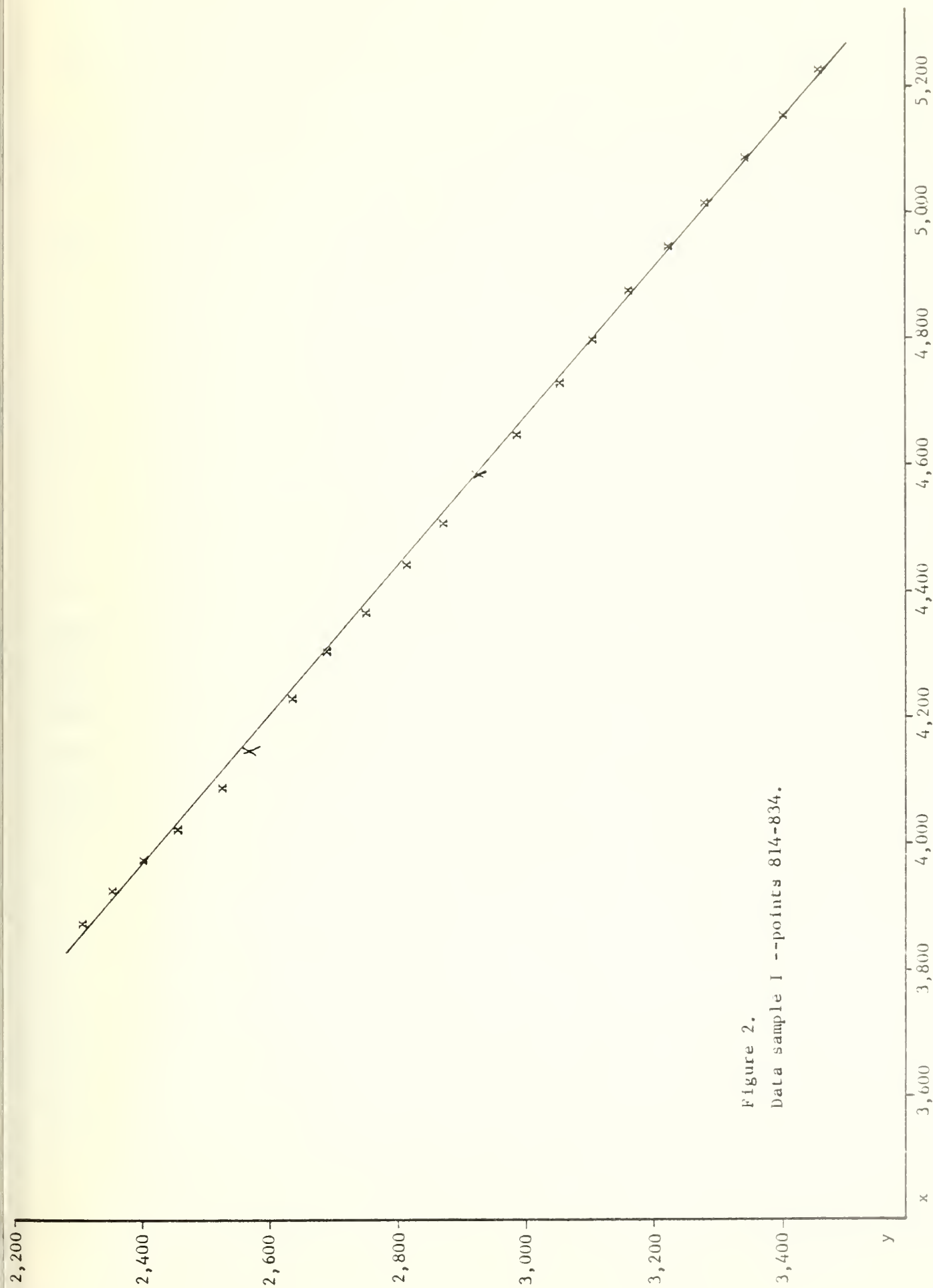


Figure 2.
Data sample I --points 814-834.

Table 3. Successive Differences — Sample I

t_i	x_i	Δ_{1i}	Δ_{2i}	Y_i	Δ_{1i}	Δ_{2i}
1	5228.6			-3465.1		
2	5156.8	-71.8	+0.1	-3407.0	+58.1	+2.7
3	5085.1	-71.7	+2.9	-3346.2	+60.8	+0.3
4	5018.3	-68.8	-5.3	-3285.1	+61.1	+1.8
5	4944.2	-74.1	+8.0	-3222.2	+62.9	-6.3
6	4878.1	-66.1	-12.0	-3165.6	+56.6	+3.2
7	4800.00	-78.1	+9.8	-3105.8	+59.8	-3.7
8	4731.7	-68.3	-11.2	-3049.7	+56.1	+6.4
9	4652.2	-79.5	+9.9	-2987.2	+62.5	-6.1
10	4583.6	-68.6	-4.3	-2930.8	+56.4	+3.9
11	4510.7	-72.9	+2.4	-2870.5	+60.2	-0.6
12	4440.2	-70.5	-2.7	-2810.8	+59.7	+1.1
13	4367.0	-73.2	+3.2	-2750.0	+60.8	-0.8
14	4297.0	-70.0	-0.9	-2690.0	+60.0	+3.3
15	4226.1	-70.9	-1.6	-2626.7	+63.3	-8.2
16	4153.6	-72.5	+2.9	-2571.6	+55.1	+4.9
17	4084.0	-69.6	+3.3	-2511.6	+69.0	+2.5
18	4017.7	-66.3	+17.2	-2449.1	+62.5	-18.2
19	3968.6	-49.1	+5.1	-2404.8	+44.3	+3.4
20	3924.6	-44.0	-12.6	-2357.1	+47.7	+0.3
21	3868.0	-56.6		-2309.1	+48.0	

Linear

$$x(t) = 5288.3 - 69.78t \quad S_{xe} = 16.73$$

$$y(t) = -3518.1 + 58.72t \quad S_{ye} = 8.33$$

Quadratic

$$x(t) = 5318.6 - 77.67t + 0.3588t^2 \quad S_{xe} = 11.62$$

$$y(t) = -3532.0 + 62.33t - 0.1642t^2 \quad S_{ye} = 6.30$$

The residual deviations:

$$e_{xi} = x_i - \hat{x}(t_i)$$

$$e_{yi} = y_i - \hat{y}(t_i)$$

are shown in Figure 3. Note that there is a definite trend in these residuals starting about time t_{18} . Note also the general trend of the residuals with a small random pattern superimposed on a curve for each residual set. Higher order polynomials could be used to remove the general curve (this was not explored). Note, further, that no violent outliers are indicated. The fitted linear function is shown in Figure 2 and the observed and predicted values for x_i and y_i are presented in Tables 4a and 4b together with the residuals in these components and the deviation

$$d_i = \sqrt{e_{xi}^2 + e_{yi}^2}$$

The sequences of signs observed in Table 4a for the e_{xi} 's and e_{yi} 's are of interest. There is a sequence of +s, followed by a sequence of -s, and ending with a sequence of +s for the e_{xi} 's. Similarly, there is a sequence of -s, followed by a sequence of +s, and ending with a sequence of -s for the e_{yi} 's. (The sign of e_{y8} can be ignored or changed since the magnitude of e_{y8} is small.)

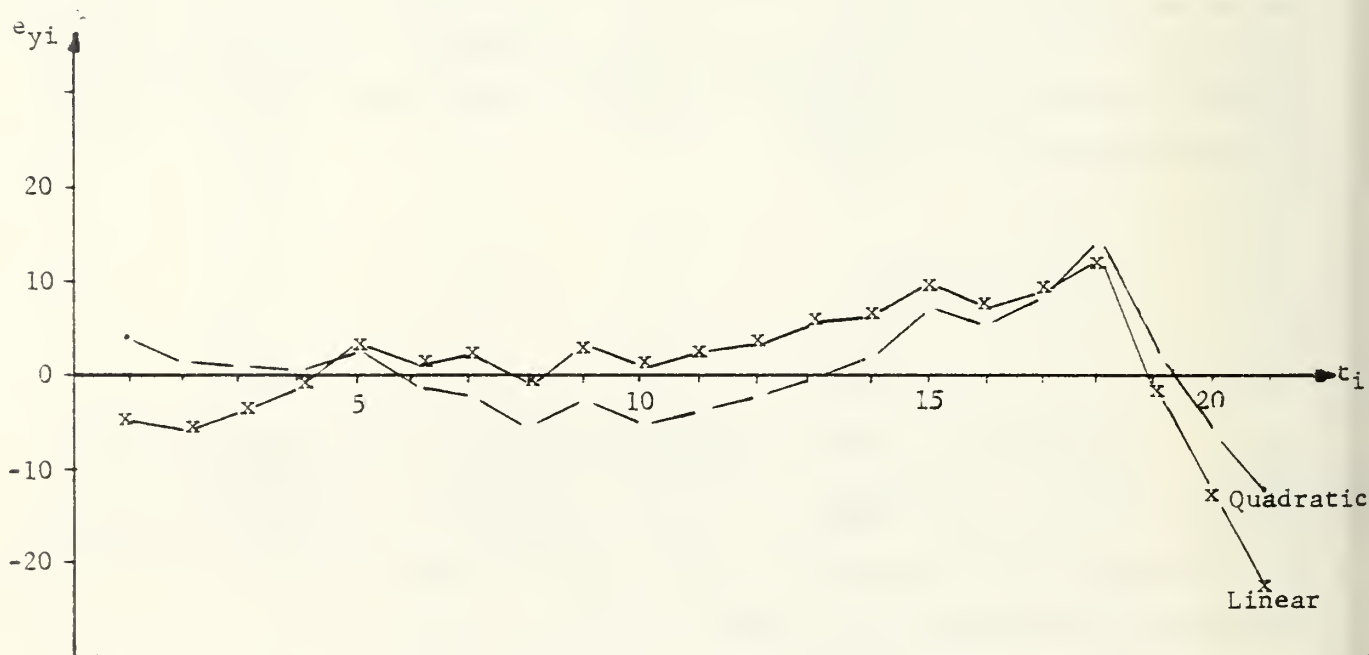
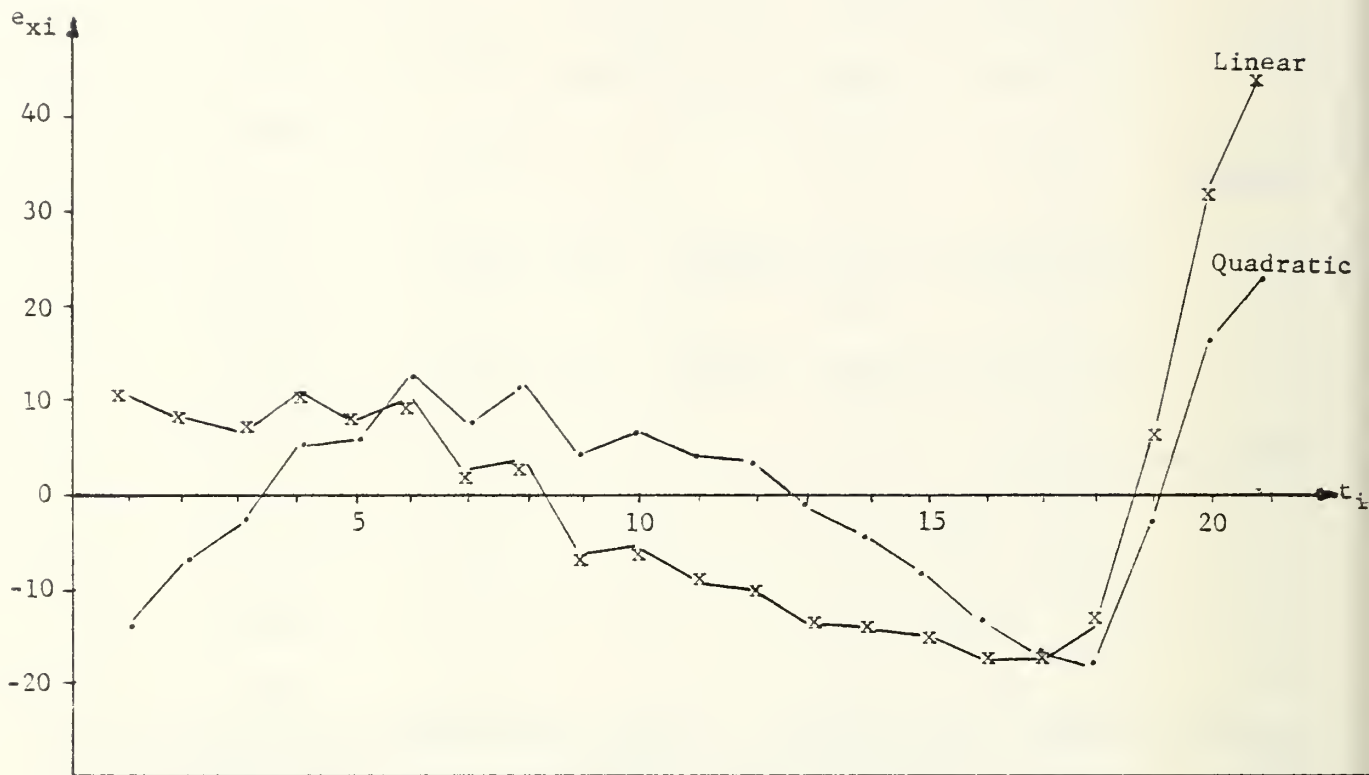


Figure 3. Least square residuals --sample I.

Table 4a. Linear Regression - Sample I

t_i	x_i	$\hat{x}(t_i)$	e_{xi}	y_i	$\hat{y}(t_i)$	e_{yi}	d_i
1	5228.6	5218.5	+10.1	-3465.1	-3459.4	-5.7	11.6
2	5156.8	5148.8	+8.0	-3407.0	-3400.7	-6.3	10.2
3	5085.1	5079.0	+6.1	-3346.2	-3342.0	-4.2	7.4
4	5018.3	5009.2	+9.1	-3285.1	-3283.2	-1.9	9.3
5	4944.2	4939.4	+4.8	-3222.2	-3224.5	+2.3	5.3
6	4878.1	4869.7	+8.4	-3165.6	-3165.8	+0.2	8.4
7	4800.0	4799.9	+0.1	-3105.8	-3107.1	+1.3	1.3
8	4371.7	4730.1	+1.6	-3049.7	-3048.4	-1.3	2.1
9	4652.2	4660.3	-8.1	-2987.2	-2989.6	+2.4	8.5
10	4583.6	4590.6	-7.0	-2930.8	-2930.9	+0.1	7.0
11	4510.7	4520.8	-10.1	-2870.5	-2872.2	+1.7	10.2
12	4440.2	4451.0	-10.8	-2810.5	-2813.5	+3.0	11.2
13	4367.0	4381.2	-14.2	-2750.0	-2754.8	+4.8	15.0
14	4297.0	4311.4	-14.4	-2690.3	-2696.0	+5.7	15.5
15	4226.1	4241.7	-15.6	-2626.7	-2637.3	+9.6	18.3
16	4153.6	4171.9	-18.3	-2571.6	-2578.6	+7.0	19.6
17	4084.0	4102.1	-18.1	-2511.6	-2519.9	+8.3	19.9
18	4017.7	4032.3	-14.6	-2449.1	-2461.2	+12.1	19.0
19	3968.6	3962.5	+6.1	-2404.8	-2402.4	-2.4	6.6
20	3924.6	3892.8	+31.8	-2357.1	-2343.7	-13.4	34.5
21	3868.0	3823.0	+45.0	-2309.1	-2285.0	-24.1	51.1

Table 4b. Quadratic Regression - Sample I

t_i	x_i	$\hat{x}(t_i)$	e_{xi}	y_i	$\hat{y}(t_i)$	e_{yi}	d_i
1	5228.6	5241.3	-12.7	-3465.1	-3469.8	+4.7	13.5
2	5156.8	5164.7	-7.9	-3467.0	-3408.0	+1.0	8.0
3	5085.1	5088.8	-3.7	-3346.2	-3346.4	+0.2	3.7
4	5018.3	5013.6	+4.7	-3285.1	-3285.3	+0.2	4.7
5	4944.2	4939.2	+5.0	-3222.2	-3224.4	+2.2	5.5
6	4878.1	4865.5	+12.6	-3165.6	-3163.9	-1.7	12.7
7	4800.0	4792.5	+7.5	-3105.8	-3103.7	-2.1	7.8
8	4731.7	4720.2	+11.5	-3049.7	-3043.8	-5.8	12.9
9	4652.2	4648.6	+3.6	-2987.2	-2984.3	-2.9	4.6
10	4583.6	4577.8	+5.8	-2930.8	-2925.1	-5.7	8.1
11	4510.7	4507.6	+3.1	-3870.5	-2866.2	-4.3	5.3
12	4440.2	4438.2	+2.0	-2810.8	-2807.6	-3.2	3.8
13	4367.0	4369.5	-2.5	-2750.0	-2749.4	-0.6	2.6
14	4297.0	4301.5	-4.5	-2690.0	-2691.5	+1.5	4.7
15	4226.1	4234.2	-8.1	-2626.7	-2633.9	+7.2	10.8
16	4153.6	4167.7	-14.1	-2571.6	-2576.7	+5.1	15.0
17	4084.0	4101.9	-17.9	-2511.6	-2519.7	+8.1	19.7
18	4017.7	4036.7	-19.0	-2449.1	-2463.2	+14.1	23.7
19	3968.6	3972.3	-3.7	-2404.8	-2406.9	+2.1	4.3
20	3924.6	3908.7	+15.9	-2357.1	-2351.0	-6.1	17.0
21	3868.0	3845.7	+22.3	-2309.1	-2295.4	-13.7	26.2

These sign sequences would ordinarily indicate that the next higher order polynomial, a quadratic, should do well in reducing the residual errors. This is not substantiated; however, as Table 4b demonstrates. The deviations in this table have four sequences of the same sign and suggest that even a cubic polynomial will not necessarily produce an excellent fit to the data—this was not explored further.

An alternative to using higher order polynomials is the reduction in sample size. This alternative was explored for the sample with $n=11$. The results are shown below:

Sample Points	Linear		Quadratic	
	S_{xe}	S_{ye}	S_{xe}	S_{ye}
814-824	3.3	2.0	—	—
819-829	2.9	1.9	2.1	1.8
824-834	16.4	9.5	—	—
829-839	13.9	11.1	—	—

The three basic causes for residuals are:

- (a) maneuver of object tracked (this is represented by the polynomial),
- (b) noise in measurements, (this is represented by σ of which S_e is an estimate), and
- (c) outliers (these will be discussed later in this report).

It is assumed that there are no outliers in Sample I. Subsample 2 (points 819 to 829) appears to be fitted quite well by a straight line and the quadratic was applied to give an estimate of the size of σ . The first subsamples (points 814 to 824) are fitted reasonably well by a straight line so the quadratic was not tried. The last two subsamples have substantially larger S_e 's. This could be caused by either torpedo maneuvers or a larger noise component (larger σ)—this was not explored.

C.

Data Sample II

The second sample selected for study was the set with times 867 to 887. These 21 points appear to present a curved path which might possibly be fitted by a quadratic. First, consider the successive differences in Table 5. Some difficulty similar to an outlier is indicated in the vicinity of $t_i = 6$ ($t_i = 872$). Examination of the first successive differences shows a drop in velocity between t_5 and t_6 and only partial recovery between t_6 and t_7 . One possible explanation would be an additional data point between t_5 and t_6 . The actual explanation is the inadvertent introduction of a measurement from a different array taken at time t_5 and entered as the measurement at t_6 . Measurements at t_7 , and subsequent times, should be shifted to respective preceding times.

Instead of fitting all of Sample II, eleven points (872-882) were selected somewhat arbitrarily for fitting by least squares—these are plotted in Figure 4. The second differences all have the same sign and the third differences are small and have apparently random sign. The least squares straight line fit is presented in Table 6a and sketched in Figure 4. (Note the shift in the time scale). This was introduced to reduce the magnitudes of the numbers calculated in determining the fitted line and S_e . In dealing with the quadratic, the means $\bar{x} = \frac{1}{11} \sum x_i$ and $\bar{y} = \frac{1}{11} \sum y_i$ were also subtracted from each observation x_i and y_i , respectively, for the same reason. Table 6b presents the quadratic regression. The reduction in the S_e 's is dramatic as would be expected from Figure 4. All of the e_i 's are less than 5 and hence within the residual noise that could be expected with a σ of 2 or 3. The signs of the e_{xi} 's; however, show some indications of lack of randomness. For this reason, a third-degree polynomial was tried for the x_i 's only. This produced the value $S_{xe} = 0.946$ with the maximum magnitude of any e_{xi} being 1.2. The cubic fits the data very well indeed.

D.

Data Sample III

The third sample selected for study involved an S-shaped maneuver as indicated by the 21 points (848-868) shown in Figure 5. The x and y coordinates of these points are presented in Figure 6 where it is evident that first and second order polynomials will not provide acceptable fits to the data. A third-order polynomial appears possible for the y_i 's and a fourth order for the x_i 's. A subset of 11 points (851-861 or points 4-14 in Figure 6 and Table 7) will be used for illustration.

Table 5. Successive Differences - Sample II

t_i	x_i	Δ_1	Δ_2	Δ_3	y_i	Δ_1	Δ_2	Δ_3
1	2012.0				-1255.5			
2	2030.0	+18.0				+94.2		
3	2056.1	+26.1	+8.1		-1161.3	+91.1	-3.1	
4	2091.0	+34.9	+8.8	+0.7	-1070.2	+88.1	-3.0	+0.1
5	2134.2	+43.2	+8.3	-0.4	-982.1	+87.0	-1.1	+1.9
6	2142.3	+8.1	-35.1	-43.4	-895.1	+20.6	-107.6	-106.5
7	2183.2	+40.9	+32.8	+70.9	-915.7	+99.7	+120.3	-227.9
8	2241.8	+58.6	+17.7	-15.1	-816.0	+99.7	-22.0	-142.3
9	2305.5	+63.7	+5.1	-12.6	-738.3	+77.7	-9.2	+12.8
10	2377.1	+71.6	+7.9	+2.8	-669.8	+68.5	-2.9	-6.3
11	2451.6	+74.5	+2.9	-5.0	-604.2	+65.6	-7.1	-7.1
12	2533.8	+85.8	+2.9	+4.8	-548.6	+55.6	-10.0	+5.2
13	2619.6	+88.1	+7.7	+4.8	-497.8	+50.8	-4.8	-2.3
14	2707.7	+91.9	+32.2	-4.1	-454.6	+43.2	-7.6	-0.8
15	2799.6	+92.0	+3.6	-1.3	-419.8	+34.8	-8.4	-0.4
16	2891.6	+95.7	+2.3	+1.5	-393.0	+26.8	-8.0	-0.4
17	2987.3	+95.9	+3.8	-3.7	-374.6	+18.4	-8.4	-2.2
18	3083.2	+94.3	+0.1	+3.6	-366.8	+7.8	-10.6	+3.5
19	3177.5	+98.7	+3.7	-3.5	-366.1	+0.7	-7.1	-4.0
20	3276.2	+93.8	+0.2	-1.8	-376.5	-10.4	-11.1	-1.7
21	3370.0		-1.6	+5.9	-399.7	-23.2	-12.8	+9.5
			+4.3	-9.2			-3.3	
			-4.9					

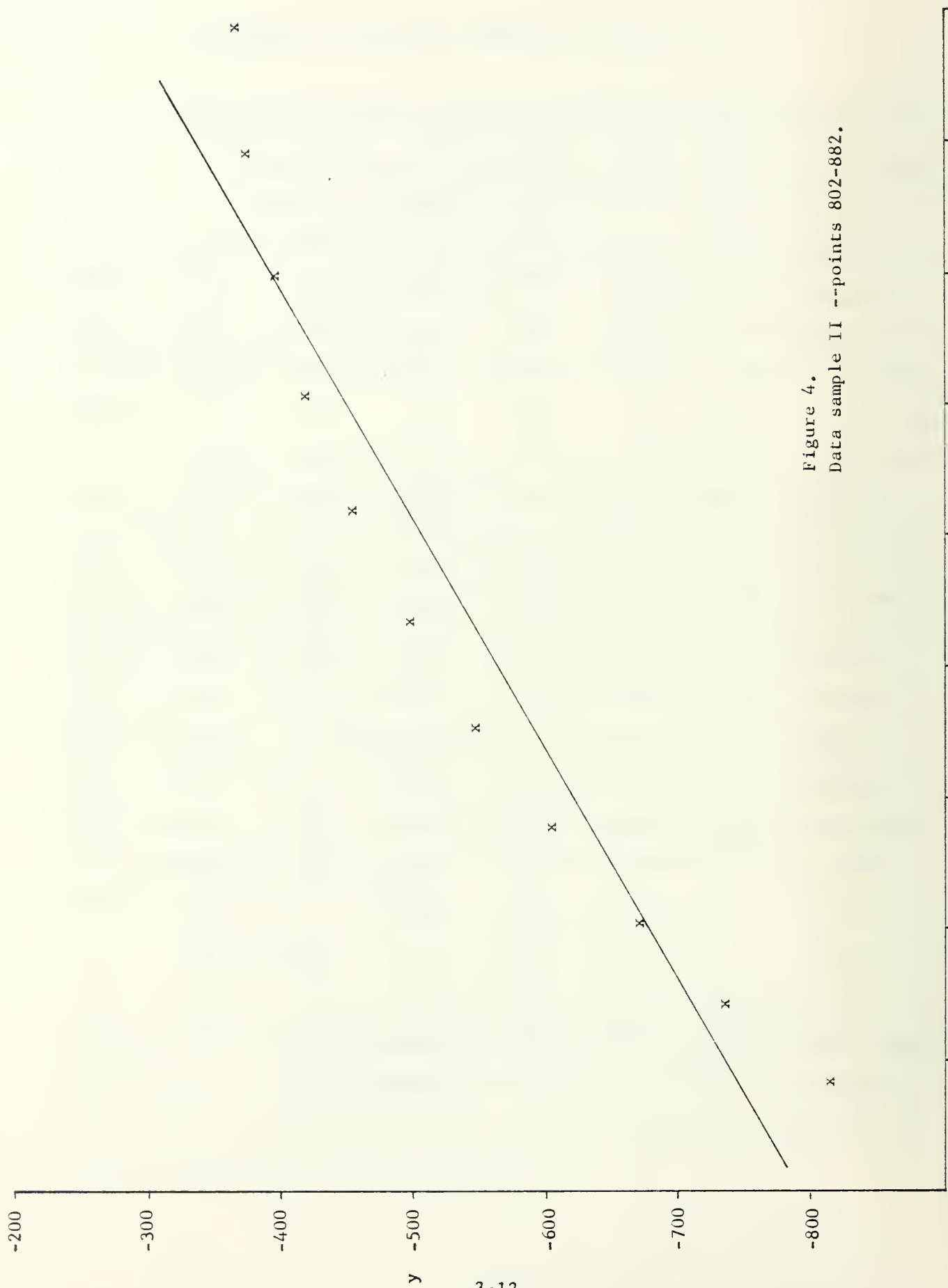


Figure 4.
Data sample II --points 802-882.

Table 6a. Linear Regression - 11 Points (872-882)

t_i	x_i	$\hat{x}(t_i)$	e_{xi}	y_i	$\hat{y}(t_i)$	e_{yi}	d_i
-5	2183.2	2148.5	+34.7	-816.0	-762.0	-54.0	64.2
-4	2241.8	2229.7	+12.1	-738.3	-716.5	-21.8	24.9
-3	2305.5	2310.9	-5.4	-669.8	-671.1	+1.3	5.6
-2	2377.1	2392.1	-15.0	-604.2	-625.7	+21.5	26.2
-1	2451.6	2473.2	-21.6	-548.6	-586.2	+31.6	38.3
0	2533.8	2554.4	-20.6	-497.8	-534.8	+37.0	42.4
1	2619.6	2635.6	-16.0	-454.6	-489.4	+34.8	38.3
2	2707.7	2716.8	-9.1	-419.8	-443.9	+24.1	25.7
3	2799.6	2798.0	+1.6	-393.0	-398.5	+5.5	5.7
4	2891.6	2879.2	+12.4	-374.6	-353.1	-21.5	24.8
5	2987.3	2960.4	+26.9	-366.1	-307.6	-58.5	64.4

$$\hat{x}(t) = 2554.4 + 81.19t$$

$$S_{xe} = 20.33$$

$$\hat{y}(t) = -534.8 + 45.43t$$

$$S_{ye} = 36.41$$

Table 6b. Quadratic Regression - 11 Points (872-882)

t_i	x_i	$\hat{x}(t_i)$	e_{xi}	y_i	$\hat{y}(t_i)$	e_{yi}	d_i
-5	2183.2	2179.3	+3.9	-816.0	-817.8	+1.8	4.3
-4	2241.8	2241.8	-0.2	-738.3	-738.9	+0.6	0.6
-3	2305.5	2308.8	-3.3	-669.8	-667.4	-2.4	4.1
-2	2377.1	2379.7	-2.6	-604.2	-603.3	-0.9	2.8
-1	2451.6	2454.7	-3.1	-548.6	-546.7	-1.9	3.6
0	2533.8	2533.9	-0.1	-497.8	-497.6	-0.2	0.2
1	2619.6	2617.1	+2.5	-454.6	-455.9	+1.3	2.8
2	2707.7	2704.5	+3.2	-419.8	-421.6	+1.8	3.7
3	2799.6	2796.0	+3.6	-393.0	-394.8	+1.8	4.0
4	2891.6	2891.6	0.0	-374.6	-375.5	-0.8	0.8
5	2987.3	2991.3	-4.0	-366.1	-363.5	-2.6	4.8

$$\hat{x}(t) = 2533.9 + 81.19t + 2.057t^2$$

$$S_{xe} = 3.32$$

$$\hat{y}(t) = -497.6 + 45.43t - 3.724t^2$$

$$S_{ye} = 1.91$$

The results of fitting third-degree polynomials to these 11 points is presented in Table 8 and the fourth-degree polynomial in Table 9. The cubic equation fits the y component quite well, but even the quartic equation leaves something to be desired (smaller S_e) for the x component. Higher order polynomials were not tried. The estimates S_e for σ obtained by fitting polynomials to the subsample of 11 points are presented below:

Order of Polynomial	X	Y
1	66.8	94.5
2	37.3	42.6
3	34.0	3.5
4	9.3	

Improvement in fitting the y component by increasing the order of the polynomial is quite dramatic but the improvement is considerably slower for the x component. The third-order polynomial could be considered acceptable for y but a fifth-order polynomial should be tried for x. The order of polynomial used does not have to be the same for both components.

E. Discussion

Only one in-water run was examined and, for it, only selected sections of the torpedo path were treated in any detail. Nevertheless some conclusions can be made about application of the Sequential Differences and Least Squares Regression techniques to 3-D data.

(1) Sequential differences:

(a) These differences provide some capability for locating isolated outlier points which differ substantially from the path of the object being tracked. This was illustrated in Sample II. The model shown in Tables 1 and 2 needs extension to higher order polynomial paths and multiple outliers. Also, the critical magnitudes for sequential differences (refer to Table 1) must be increased to allow for accelerations since the use of sequential differences will precede fitting a polynomial and hence the order of the fitted polynomial will not be known at the time. Thus sequential differences should be used only for a first screening for gross outliers.

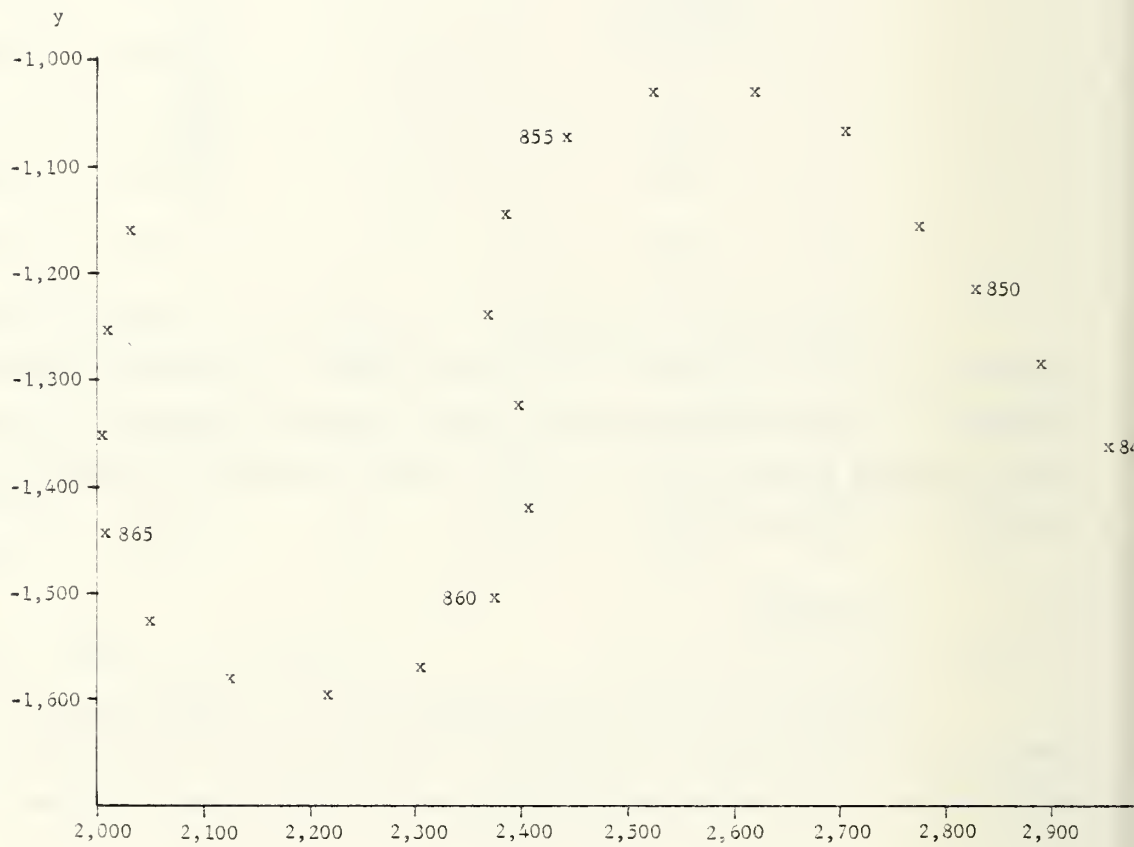


Figure 5. Data sample III --points 848-868.

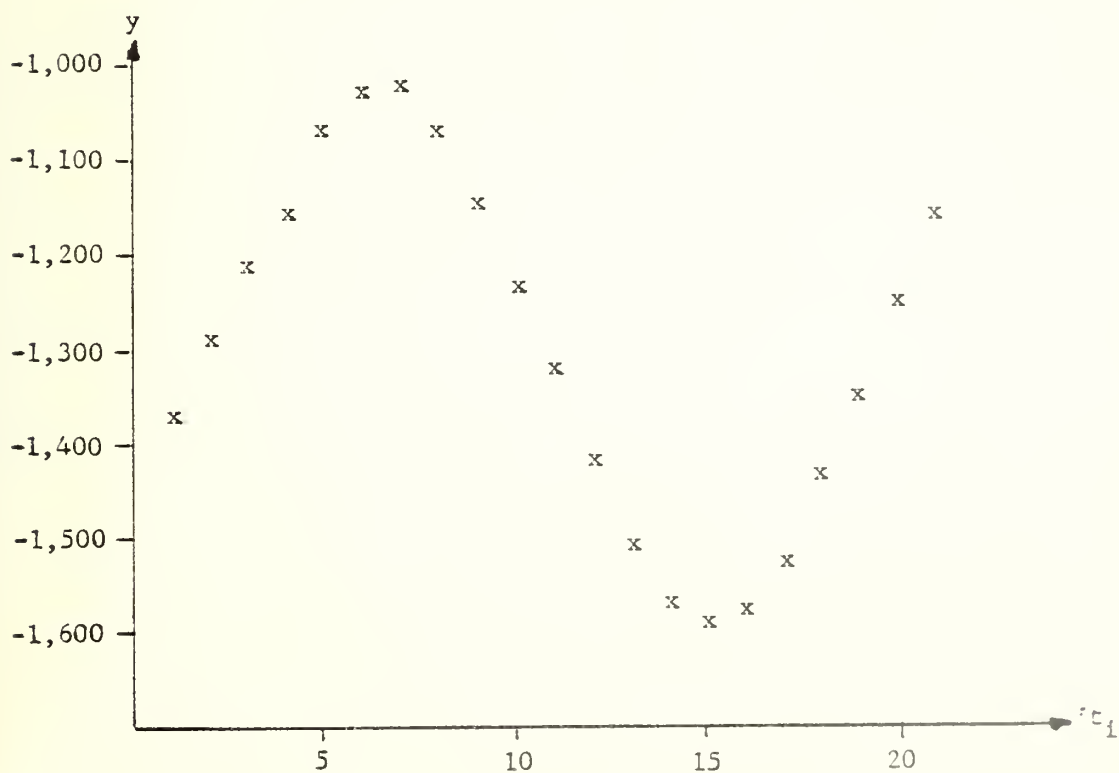
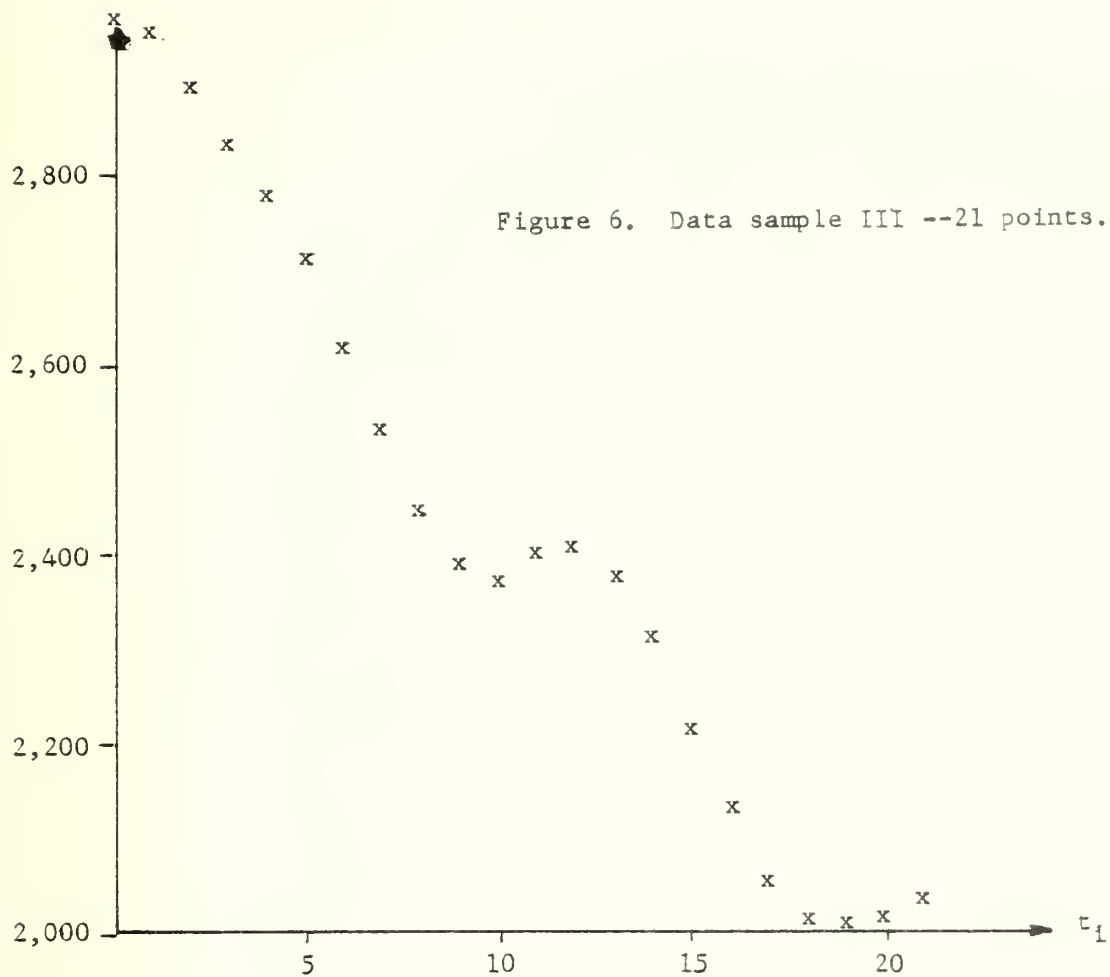


Table 7. Successive Differences - Sample III

t_i	x_i	Δ_1	Δ_2	Δ_3	y_i	Δ_1	Δ_2	Δ_3
1	2949.3				-1364.0			
2	2889.3	-40.5				+74.4		
3	2828.5	-56.8	-10.8		-1289.6	+74.0	-0.4	
4	2777.0	-51.5	-0.7	+10.1	-1215.6	+56.5	-17.5	-17.9
5	2702.5	-74.5	-23.0	-22.3	-1159.1	+88.8	+32.3	+49.8
6	2617.8	-84.7	-10.2	+12.8	-1070.3	+37.5	-51.3	-83.6
7	2524.3	-93.5	-8.8	+1.4	-1032.8	+1.0	-36.5	+14.8
8	2440.0	-84.3	+9.2	+18.0	-1031.8	-44.0	-45.0	-8.5
9	2385.7	-54.3	+30.0	+20.8	-1075.8	-72.4	-28.4	+16.6
10	2369.8	+25.7	+38.4	+8.4	-1148.2	-91.7	-19.3	+9.1
11	2395.5	+10.6	+41.6	+3.2	-1239.9	-78.9	+12.8	+32.1
12	2406.1	-33.0	-15.1	-56.7	-1328.8	-91.4	-12.5	-25.3
13	2373.1	-67.8	-43.6	-28.5	-1420.2	-88.5	+2.9	+15.4
14	2305.3	-89.2	-34.8	+8.8	-1508.7	-64.4	+24.1	+21.2
15	2216.1	-91.4	-21.4	+13.4	-1573.1	+39.4	+15.3	
16	2124.7	-75.7	-2.2	+19.2	-1598.1	+16.8	+2.4	
17	2049.0	-42.2	+15.7	+17.9	-1581.3	+53.7	+41.8	-4.9
18	2006.8	-4.5	+33.5	+17.8	-1527.6	+83.6	+36.9	-7.0
19	2002.3	+9.7	+37.7	+4.2	-1440.0	+93.2	+29.9	-20.3
20	2012.0	+18.0	+14.2	-23.5	-1350.8	+95.3	+9.6	-7.5
21	2030.0		+8.3	-5.9	-1255.5	+94.2	+2.1	-3.2
					-1161.3		-1.1	

Table 8. Cubic Regression - Sample III (11 points)

t_i	x_i	$\hat{x}(t_i)$	e_{xi}	y_i	$\hat{y}(t_i)$	e_{yi}	d_i
-5	2777.0	2804.8	-27.8	-1059.1	-1159.0	-0.1	27.8
-4	2702.5	2680.2	+22.3	-1070.3	-1066.7	-3.6	22.6
-3	2617.8	2383.9	+33.9	-1032.8	-1028.6	-4.2	34.2
-2	2524.3	2511.8	+12.5	-1031.8	-1035.5	+3.7	13.0
-1	2440.0	2459.6	-19.6	-1075.8	-1078.3	+2.5	19.8
0	2385.7	2423.3	-37.6	-1148.2	-1147.7	-0.5	37.6
1	2369.8	2398.6	-28.8	-1239.9	-1234.7	-5.2	29.3
2	2395.5	2381.3	+14.2	-1328.8	-1330.0	+1.2	14.3
3	2406.1	2637.6	+38.5	-1420.2	-1424.5	+4.3	38.7
4	2373.1	2252.8	+20.3	-1508.7	-1509.1	+0.4	20.3
5	2305.3	2333.1	-27.8	-1573.1	-1574.5	+1.4	27.8

$$\hat{x}(t) = 2423.3 - 29.812t + 5.827t^2 - .649308t^3$$

$$\hat{y}(t) = -1147.7 - 79.73t - 8.761t^2 + 1.5271t^3$$

$$S_{xe} = 34.0 \quad S_{ye} = 3.5$$

Table 9. Quartic Regression - Sample III (11 points)

t_i	x_i	$\hat{x}(t_i)$	e_{xi}
-5	2777.0	2774.0	+3.0
-4	2702.5	2711.0	-8.5
-3	2617.8	2614.8	+3.0
-2	2524.3	2516.9	+7.4
-1	2440.0	2439.1	+0.9
0	2385.7	2392.5	-6.8
1	2369.8	2378.1	-8.3
2	2395.5	2386.6	+8.9
3	2406.1	2398.4	+7.7
4	2373.2	2383.7	-10.6
5	2305.3	2302.3	+3.0

$$\hat{x}(t) = 2392.4 - 29.812t + 16.533t^2 - .6943t^3 - .428234t^4$$

$$S_{xe} = 9.3$$

(b) Sequential differences also provide some indication of the order of polynomial that will be required. One indicator is the number of sign changes that occur on the successive differences of a particular order. If there are few sign changes, then a non-random effect is indicated and a higher order polynomial will be indicated. Thus, for example, in Sample II the 11-point data subset shows a long sequence of +'s for the Δ_{2i} 's, but no such sequence (indicating randomness) for the Δ_{3i} 's. Hence, a third order polynomial can be expected to provide some improvement over a second-order polynomial. This type of information may be difficult to incorporate into a data smoothing algorithm, but even some simple procedure can be of help in reducing the computational load.

(2) Sample Size:

(a) Although it is possible that a sample of 21 points could be fitted with acceptably small S_e in some instances (the quadratic was not tried on Sample II), it would appear that smaller samples (e.g., $n=11$) will allow fitting the data with a reasonably low-order polynomial. The size $n=11$ is not sacrosanct but will leave some room for elimination of outliers and so seems to be a reasonable size.

(3) Least squares smoothing:

(a) By its nature, the estimate S_e , for the standard deviation σ of the measurement noise, is monotone decreasing as the order of the polynomial increases. (An $n-1$ order polynomial should be able to fit n points exactly so that S_e would be zero.) The appropriate order polynomial is one which reduces S_e to the level of the noise in the measurements. This may vary with the path and the array making the measurements. For the portions of the path examined, it is suspected that σ_v is less than σ_x since S_{ye} is generally smaller than S_{xe} for a given order polynomial. The decision to use a higher-order polynomial to fit a set of data depends upon the value of S_e obtained for a given-order polynomial. If S_e is small (3 or 4), then higher-order polynomials cannot be expected to give much improvement. The extent to which S_e can be reduced will depend upon the component as well as the polynomial degree.

(4) Outliers:

(a) In addition to rough screening for outliers by sequential differences, there is additional screening that can be performed using residual errors after a polynomial has been fitted to the data. Outliers contributed substantially to S_e and the two basic techniques of reducing S_e are elimination of points with large residuals, or increasing the order of the polynomial.

(b) Elimination of outliers using residuals after smoothing can be accomplished in two ways:

(1) by confidence intervals—a residual greater in magnitude than some specified multiple (3 or larger) of S_e can be considered to be a outliers, and

(2) by variance reduction—the ratio of S_e 's before and after removal of a point, or points, with substantial residuals can be used as a basis for the decision on whether to remove the points. For example, if S_e (after)/ S_e (before) $\leq r$, then the points should be removed (Grubbs' criteria). The value of r is in the range 0.0 to 1.0 and could be changed depending upon the magnitude of S_e .

(5) Sampling rate:

(a) The smoothing of 3-D data can be performed to provide either a parametric representation of path segments, or specific information such as position and velocity information, only at certain points on the path. These will be called "parametric estimation" and "point estimation," respectively.

(b) To illustrate parametric estimation, consider data collected at 200 sequential observation times (e.g., 800 to 1,000 for the 3-D data used in this section). Samples of 11 points will be used. Sample S_1 will consist of points 1 through 11, sample S_2 of points 10 through 20 and, in general, sample S_j of points from $10(j-1)$ to $10j$. There will then be 20 samples on the path. Each sample of 11 points is to be fitted by a polynomial of appropriate degree and the parameters of the polynomial together with the value of S_e recorded for the path segment represented by that sample. Note that there will be two points of overlap between S_1 and S_2 and one point of overlap thereafter.

(c) For point estimation, sequence of points must be provided. For data consisting of 200 points it may be considered that occasional monitoring is sufficient for points 0 to 50 and 100 to 150, but that behavior of the path from points 50 to 100 should be monitored more often and behavior from points 150 to 200 should be followed closely. Then the following sequence of points could be considered reasonable:

<u>j</u>	<u>Points in S_j</u>	<u>Midpoint t_i</u>
1	5-15	10
2	25-35	30
3	45-55	50
4	55-65	60
5	65-75	70
6	75-85	80
7	85-95	90
8	95-105	100
9	115-125	120
10	135-145	140
11	145-155	150
12	150-160	155
13	155-165	160
14	160-170	165
15	165-175	170
16	170-180	175
17	175-185	180
18	180-190	185
19	185-195	190
20	190-200	195

(d) At each midpoint time t_j , the position coordinate estimates, the velocity in these components, the resultant velocity, and S_{ej} can be recorded together with additional information, such as acceleration components, if desired. Note that the sequence of 20 points suggested above has substantial overlap of samples in some cases and data gaps between samples in other cases. This was introduced intentionally since least squares smoothing produces better estimates (smaller confidence intervals) at the midpoint of the sample when the fitted curve is a straight line (refer to Appendix B).

(e) Parametric estimation could also be modified to delete some samples (e.g., alternate samples from $t_j=100$ to $t_j=150$). It should require greater modification to achieve the quality of point estimation procedure at other than parametric sample midpoints when a straight line (first-order polynomial) is used. When higher order polynomials are required, the preference for the best estimate at midpoint of the sample is lost (refer to Appendix B). Making both techniques available provides some flexibility in data smoothing to accomodate potential customers.

IV. A DATA SMOOTHING ALGORITHM

The following procedure is suggested for smoothing 3-D data:

Step 1: Select appropriate sample size. (11 is suggested as being small enough to provide some capability of fitting path segments of maneuvering torpedoes without requiring high-order polynomials. Some leeway for dropping outliers is also provided.)

Step 2: Select parameter of point estimation.

Step 3: Select sampling rate. (A standard rate such as described in Section III E4 should be provided as a default rate for parameter estimation and the midpoints of these samples as a default rate for point estimation.)

Step 4: Adjust data for missing data points. (The principle applied here is minimization of the effect of the numbers on sequential differences. For a single missing datum, the average of the values at two adjacent times will minimize the second differences. In any case, data supplied in this step must be removed before least squares smoothing is applied.)

Step 5: Calculate first, second, and third order sequential differences.

Step 6: Determine approximate polynomial order k . (The $(k+1)^{\text{st}}$ order sequential differences should contain noise only, and thus, have random signs. Sequences of 4, or more, differences with the same sign suggest the presence of a non-random component as does the occurrence of 4, or fewer, changes of sign. The presence of a non-random component is going to be awkward to identify. If the second differences are random, then $k=1$. If the second-order differences are non-random, but the third-order differences are random, then $k=2$. If the third-order differences are non-random then fourth-order differences should be calculated and examined for randomness. (This examination of sequential differences in increasing order should probably not be carried beyond the fifth-order.)

Step 7: Screen successive differences for gross outliers. (This must follow determination of approximate degree of polynomial since it should be based on comparison of magnitude of deviation to noise only as indicated in Tables 1 and 2. The critical values suggested in those tables should be increased substantially. Some limit, possibly between 50 and 100, should be selected keeping in mind that this is a first screening for gross outliers and a second screening will be made. Any outliers found in this step; however, will reduce computations in later steps. Remove any outliers found and the observations for the other space components at the same observation time.)

Step 8: Check for polynomial degree compatibility. (If the number of outliers removed (r) satisfies the inequality $r + k \geq n - 1$, where k is the degree of polynomial found in Step 6 and n is the sample size after data points supplied in Step 4 are removed, then fitting a k^{th} order polynomial will be inappropriate. For example, if $r = 4$ points are removed from a sample in which one data point has been created in Step 4, then a polynomial of degree 5 can be fitted to the data without any residual errors since there are 6 linear relationships of the 6 coefficients.)

Step 9: Fit a polynomial of degree k to the data. (The least squares procedure outlined in Appendix A is applicable. At this step only S_{ke} need be determined and not the coefficients.)

Step 10: Seek acceptable S_e . (If S_{ke} is unacceptably large, repeat Step 9 with k replaced by $k + 1$. Repeat this step until either S_e is acceptable or a polynomial of degree 5 is fitted to the data.)

Step 11: Complete least squares polynomial fit. (The coefficients for the polynomial of degree found in Step 10 are now needed, and the residual errors.)

Step 12: Second screening for outliers. (One of the procedures discussed in Section III E3 should be applied to locate any outliers not found in Step 7. Remove the outliers).

Step 13: Repeat Steps 9, 10, 11, and 12 until no more outliers are found. (The polynomial obtained will be used for smoothing sample data. Note that the alternative procedure of searching residuals for each polynomial degree to locate outliers may result in removing points which are not actually outliers but legitimate observations for a higher

degree polynomial. On the other hand, the proposed method could use a higher order polynomial to fit outliers when a lower order polynomial should actually be used. There is a choice of the type of misfit that is acceptable.)

Step 14: Record smoothed path. (For parametric form, if specified in Step 2, recorded data includes coefficients of fitted polynomial, S_{ej} and n_j for each sample S_j specified in Step 3. For point estimation form, if specified in Step 2, recorded data includes: time t_j , estimated coordinates $\hat{s}_j = x(t_j)$, $\hat{y}_j = y(t_j)$, and $\hat{z}_j = z(t_j)$, velocity components, S_{ej} , and n_j for each point specified in Step 3. Additional path information may also be specified; e.g., acceleration components.)

V. CONCLUSIONS AND RECOMMENDATIONS

The procedure suggested in Section IV provides a reasonable approach for obtaining the information desired in parts (1), (2), and (3) of Section I B. No attempt has been made to provide the information in part (4).

In instrumenting this procedure, several parameters must be provided:

A. Sample Size (Step 1)

A smaller sample size of $n=7$ has been suggested. This would permit fitting path segments contained maneuvers with lower order polynomials, but is subject to greater degradation by missing data points and removal of outliers. Experience on relative occurrence of such events in actual field data will be useful in selecting appropriate sample size.

B. Choice of Parameter or Point Estimation (Step 2) and Sampling Rate (Step 3)

The desires of the customers who will use the smoothed data is of primary concern here.

C. Specifying Approximate Polynomial Order (Step 6)

It will be difficult to specify a simple rule for determining that the k^{th} order sequential differences contain non-random components but the $(k+1)^{\text{st}}$ order differences involve only random components. The Theory of Runs can be of some help here although a simpler rule is desirable—this needs further study.

D. Rough Screening For Outliers (Step 7)

A reasonable critical level for identifying outliers by sequential differences must be established. The occurrence of an isolated outlier was considered in Section II B. Other potential producers of large sequential differences such as paired outliers, violent changes in velocity, et cetera, should be examined for resultant effects. Identification of signatures for such effects will be useful in using sequential differences to identify outliers.

E. Polynomial Degree Limitations (Step 10).

The limitation of polynomial degree to 5, or less, appears reasonable for samples of size 11. The possibility of decreasing this limit to 4 or increasing it to 6 or higher should be considered. This may require more experience with in-water run data. For smaller sample sizes, such as $n=7$, reduction of this limit to lower polynomial degree should be considered.

F. Computing Smoothed Path (Step 11)

The pivotal condensation method outlined in Appendix A can be simplified even further in certain cases which may occur frequently enough to take advantage of their commonality in the computer program. In particular, when the sample consists of $n=11$ data points at adjacent times, the shift of the time origin to the midpoint of the sample produces the following effects:

(1) coefficients of the polynomial parameters are the same in the normal equations for all samples,

(2) only the last column in the pivotal condensation format changes with sample, and

(3) the other columns in the pivotal condensation format require only addition of a row and a column in each box when the next higher degree polynomial is considered.

The above commonality is also clearly evident in the vector representation presented in Appendix A. The extent to which this commonality can be exploited depends primarily upon the rarity of missing data points and outliers. Indeed, depending upon requirements of the ultimate users, data smoothing could conceivably be restricted to only such samples.

In summary, the data smoothing algorithm presented in Section IV appears reasonable, but there are several elements that must be specified before it can be implemented. Some of these can be improved by further research, others depend upon the quality of the data which can only be determined by experience with actual 3-D data. Finally, some of them can only be determined in consultation with the ultimate users of the smoothed data.

A P P E N D I X A

L E A S T S Q U A R E S D A T A S M O O T H I N G

A-1 LINEAR LEAST SQUARES WITH ONE PREDICTOR

Sample:

$$(x_i \ y_i) \quad i=1,2,\dots,n$$

Assumptions:

is linear, i.e., A1 -- Actual relationship between X and Y

$$\tilde{y}(\tilde{x}) = \alpha + \beta \tilde{x}$$

 A2 -- Abscissas are without errors

$$x_i = \tilde{x}_i$$

 A3 -- Ordinates contain measurement or
observations/errors

$$y_i = \tilde{y}_i + \epsilon_i$$

ϵ_i = observational error

$$\tilde{y}_i = \tilde{y}(x_i)$$

Problems:

Fit a straight line to the data

Engineer's Solution:

$$\text{Let} \quad \hat{y}(x) = a + bx$$

$$e_i = y_i - \hat{y}(x_i)$$

$$D = \sum_{i=1}^n e_i^2 = \sum_{i=1}^n (y_i - a - bx_i)^2$$

The coefficients a and b are selected to minimize D (the sum of squares of the deviations of the observed y_i 's from the fitted line). Setting

$$\frac{\partial D}{\partial a} = 0 \quad \text{and} \quad \frac{\partial D}{\partial b} = 0$$

gives the two equations

$$na + (\sum x_i) b = \sum y_i$$

$$(\sum x_i) a + (\sum x_i^2) b = \sum x_i y_i$$

Solving these equations yields the desired estimates a and b for the parameters α and β , i.e.,

$$b = \frac{n(\sum x_i y_i) - (\sum x_i)(\sum y_i)}{n(\sum x_i^2) - (\sum x_i)^2}$$

$$a = \frac{(\sum y_i) - b(\sum x_i)}{n}$$

Computational Format:

The following format uses pivotal condensation to produce a and b. It also yields D and hence the sample variance

$$S_e^2 = \frac{1}{n-2} \sum_{i=1}^n e_i^2$$

without requiring calculation of the individual e_i 's.

n	$(\sum x_i)$	$(\sum y_i)$	$A_{xx} = [n(\sum x_i^2) - (\sum x_i)^2] / n$
	$(\sum x_i^2)$	$(\sum x_i y_i)$	$A_{xy} = [n(\sum x_i y_i) - (\sum x_i)(\sum y_i)] / n$
		$(\sum y_i^2)$	$A_{yy} = [n(\sum y_i^2) - (\sum y_i)^2] / n$
	A_{xx}	A_{xy}	$D = (A_{xx} A_{yy} - A_{xy}^2) / A_{xx}$
		A_{yy}	$S_e^2 = D / (n-2)$
		D	$b = A_{xy} / A_{xx}$
a	b	S_e^2	$a = [(\sum y_i) - b(\sum x_i)] / n$

Statistician's Solution:

Statisticians augment the Engineer's Solution by adding the following assumptions:

A4 -- The observational errors (the ϵ_i 's) are realizations of independent random variables, E_i 's, with zero means and common variance, i.e.,

$$\mu_{E_i} = \mathcal{E}(E_i) = 0$$

$$\sigma_{E_i}^2 = \mathcal{E}(E_i - \mu_{E_i})^2 = \sigma^2 \text{ for all } i.$$

A5 -- The observational errors are normally distributed random variables. This will be denoted by

$$E_i \sim N(0, \sigma^2)$$

Now let \tilde{y}_i denote the realization of the random variable Y_i . Then

$$Y_i = \tilde{y}(X_i) + E_i$$

and

$$\hat{y}(Y_i) = \tilde{y}(x_i)$$

Further, the random variable \hat{Y}_i can be expressed in the form

$$\hat{Y}_i = A + BX_i$$

where

$$B = \frac{n \sum x_i Y_i - (\sum x_i)(\sum Y_i)}{n(\sum x_i^2) - (\sum x_i)^2} = \frac{A_{xy}}{A_{xx}} \quad \text{and}$$

$$A = \frac{(\sum Y_i) - B(\sum x_i)}{n}$$

Note that A and B are linear functions of the Y_i 's and hence of the E_i 's. It can now be shown that

$$\mu_A = \hat{y}(A) = \alpha \quad \text{and}$$

$$\mu_B = \hat{y}(B) = \beta$$

so that a and b are unbiased estimators for α and β . The evaluation of the variances of $\hat{y}(x)$, A and B is simplified if the x_i 's are shifted so that their mean is zero. Then, since

$$\begin{aligned} \sum x_i &= 0 \\ A_{xx} &= \sum x_i^2 \\ A_{xy} &= \sum x_i Y_i \\ b &= \frac{(\sum x_i Y_i)}{(\sum x_i^2)} \\ a &= \frac{\sum Y_i}{n} = \bar{Y} \end{aligned}$$

This shift in the x-axis will be assumed in the development which follows.

It can now be demonstrated that

$$\begin{aligned}\sigma_{\tilde{y}_i}^2 &= \sigma^2, \\ \sigma_A^2 &= \sigma^2/n, \\ \sigma_B^2 &= \sigma^2 / (\sum x_i^2), \\ \text{Cov}(A, B) &= E[(A - \alpha)(B - \beta)] = 0, \\ \sigma_{\hat{y}(x)}^2 &= \left[\frac{1}{n} + \frac{x^2}{(\sum x_i^2)} \right] \sigma^2, \quad \text{and} \\ E(S_E^2) &= \sigma^2.\end{aligned}$$

The last relationship is very important since S_E^2 is an unbiased estimator of σ^2 and is our only source of information on this parameter.

The assumption of normality (A5) together with linearity of the other random variables in the E_i 's insures that they are also normally distributed, i.e.,

$$\begin{aligned}Y &\sim N(\tilde{Y}, \sigma^2), \\ A &\sim N(\alpha, \sigma^2/n), \\ B &\sim N(\beta, \sigma^2 / \sum x_i^2), \quad \text{and} \\ \hat{Y}(x) &\sim N[\tilde{Y}(x), (\frac{1}{n} + \frac{x^2}{\sum x_i^2}) \sigma^2]\end{aligned}$$

The random variable $(n-2) S_E^2 / \sigma^2$ has a Chi-Square distribution with $n-2$ degrees of freedom and the random variables.

$$\begin{aligned}T_\alpha &= \frac{\sqrt{n}(A - \alpha)}{S_E} \\ T_\beta &= \frac{B - \beta}{S_E \sqrt{\sum x_i^2}} \quad \text{and} \\ T_{\hat{y}}[x] &= \frac{(\hat{Y}(x) - \tilde{Y}(x))}{S_E \sqrt{\frac{1}{n} + \frac{x^2}{\sum x_i^2}}}\end{aligned}$$

have a Student-T distribution with $n-2$ degrees of freedom.

These distributions can then be used to establish confidence intervals for α , β , and $\tilde{y}(x)$ at any x . Thus, for example, with k from Student-T tables such that

$$P(-k \leq T \leq +k) = .95$$

we have the following 95% confidence intervals

$$(a - \frac{kS_e}{\sqrt{n}}, a + \frac{kS_e}{\sqrt{n}})$$

for α ,

$$(b - kS_e \sqrt{\sum x_i^2}, b + kS_e \sqrt{\sum x_i^2})$$

for β , and

$$(\hat{Y}(x) - kS_e \sqrt{\frac{1}{n} + \frac{x^2}{\sum x_i^2}}, \hat{Y}(x) + kS_e \sqrt{\frac{1}{n} + \frac{x^2}{\sum x_i^2}})$$

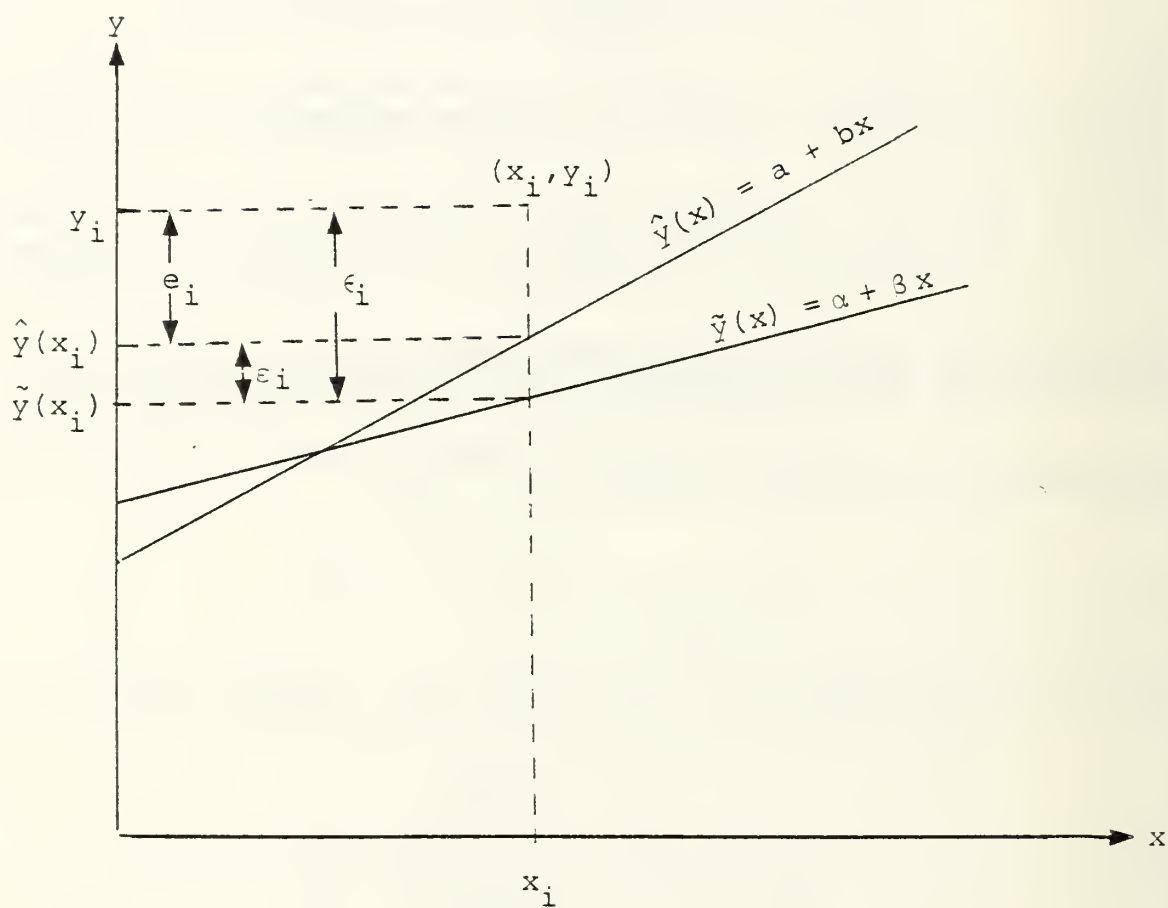
for $\tilde{y}(x)$ at any x . It should be stressed that the confidence interval for $\tilde{y}(x)$ given above involves measurements about the mean of x ($\bar{x}=0$). The general form for this confidence interval is

$$(\hat{Y}(x) - kS_e \sqrt{\frac{1}{n} + \frac{(x - \bar{x})^2}{\sum (x_i - \bar{x})^2}}, \hat{Y}(x) + kS_e \sqrt{\frac{1}{n} + \frac{(x - \bar{x})^2}{\sum (x_i - \bar{x})^2}})$$

It should be noted that the confidence interval for $\tilde{y}(x)$ is shortest for $x=\bar{x}$ and increases as x deviates from this value.

A sketch of the situation can help clarify the mathematical elements involved.

$\tilde{y}\{x\} = \alpha + \beta x$ = actual linear relationship
 $\hat{y}\{x\} = a + bx$ = fitted
 ϵ_i = observational error
 e_i = fitting error
 $\epsilon(x)$ = prediction error at any x



Sample:

$$(x_{1i}, x_{2i}, y_i) \quad i = 1, 2, \dots, n$$

Assumptions:

$$A1 \quad \tilde{y}(\tilde{x}) = \alpha_0 + \alpha_1 \tilde{x}_1 + \alpha_2 \tilde{x}_2 \quad x = (x_1, x_2)$$

$$A2 \quad x_{1j} = \tilde{x}_{1i} \quad \text{and} \quad x_{2i} = \tilde{x}_{2i}$$

$$A3 \quad y_i = \tilde{y}(x_i) + \epsilon_i \quad x_i = (x_{1i}, x_{2i})$$

Engineers' Solution:

$$\text{Let} \quad \hat{y}(x) = a_0 + a_1 x_1 + a_2 x_2$$

$$e_i = y_i - \hat{y}(x_i)$$

$$D = \sum e_i^2 = \sum (y_i - a_0 - a_1 x_{1i} - a_2 x_{2i})^2$$

Minimizing D

$$\frac{\partial D}{\partial a_0} = 0, \quad \frac{\partial D}{\partial a_1} = 0, \quad \frac{\partial D}{\partial a_2} = 0$$

produces the normal equations

$$na_0 + (\sum x_{1i})a_1 + (\sum x_{2i})a_2 = (\sum y_i) \quad (1)$$

$$(\sum x_{1i})a_0 + (\sum x_{1i}^2)a_1 + (\sum x_{1i}x_{2i})a_2 = (\sum x_{1i}y_i) \quad (2)$$

$$(\sum x_{2i})a_0 + (\sum x_{1i}x_{2i})a_1 + (\sum x_{2i}^2)a_2 = (\sum x_{2i}y_i) \quad (3)$$

which can be solved for a_0 , a_1 , and a_2 in terms of sample data.

Solving (1) for a_0 gives

$$a_0 = [(\sum y_i) - (\sum x_{1i})a_1 - (\sum x_{2i})a_2] / n \quad (1')$$

Substituting (1') in (2) and (3) gives

$$A_{11}a_1 + A_{12}a_2 = A_{1y} \quad (2')$$

$$A_{12}a_1 + A_{22}a_2 = A_{2y} \quad (3')$$

Solving (2') for a_1 gives

$$a_1 = (A_{14} - a_2 A_{12}) / A_{11}$$

and substituting in (3') gives

$$a_2 = \frac{B_{2Y}}{B_{22}}$$

where the coefficients will be defined in the computational format which follows. Equations (3''), (2'') and (1') can be used to determine the values of a_0 , a_1 , and a_2

COMPUTATIONAL FORMAT

n	Σx_{1i}	Σx_{2i}	ΣY_i	$A_{jk} = [\eta(\Sigma x_{ji} x_{ki}) - (\Sigma x_{ji})(\Sigma x_{ki})] / n$
	Σx_{1i}^2	$\Sigma x_{1i} x_{2i}$	$\Sigma x_{2i} Y_i$	$A_{jY} = [\eta(\Sigma x_{ji} Y_i) - (\Sigma x_{ji})(\Sigma Y_i)] / n$
		Σx_{2i}^2	$\Sigma x_{2i} Y_i$	$A_{YY} = [\eta(\Sigma Y_i^2) - (\Sigma Y_i)^2] / n$
			ΣY_i^2	$B_{22} = [A_{11} A_{22} - A_{12}^2] / A_{11}$
	A_{11}	A_{12}	A_{1Y}	$B_{2Y} = [A_{11} A_{2Y} - A_{12} A_{1Y}] / A_{11}$
		A_{22}	A_{2Y}	$B_{YY} = [A_{11} A_{YY} - A_{1Y}^2] / A_{11}$
			A_{YY}	$De = [B_{22} B_{YY} - B_{2Y}^2] / B_{22}$
		B_{22}	B_{2Y}	
			B_{YY}	
			De	
a_0	a_1	a_2	S_e^2	$S_e^2 = \frac{1}{n-3} \sum e_i^2 = De / (n-3)$
				$a_2 = B_{2Y} / B_{22}$
				$a_1 = [A_{1Y} - a_2 A_{12}] / A_{11}$
				$a_0 = [(\Sigma Y_i) - a_2 (\Sigma x_{2i}) - a_1 (\Sigma x_{1i})] / n$

Statisticians' Solution

Assumptions A4 and A5 lead to the following random variables and their distributions

E = observational error in y at (x_1, x_2)

$$\sim N(0, \sigma^2)$$

$$Y(x_1, x_2) = \tilde{Y}(x_1, x_2) + E$$

$$\sim N(\tilde{Y}(x_1, x_2), \sigma^2)$$

$$A_2 = B_{2Y} / B_{22}$$

$$\sim N(a_2, \frac{A_{11}}{A_{11} A_{22} - A_{12}^2} \sigma^2)$$

$$A_1 = [A_{1Y} - A_{21} A_{12}] / A_{11}$$

$$\sim N \left(\alpha_1, \frac{A_{22}}{\sqrt{A_{11} A_{11} - A_{12}^2}} \sigma^2 \right)$$

$$A_0 = [(\sum Y_i) - A_1 \sum x_{1i} - A_2 \sum x_{2i}] / n$$

$$\sim N(\alpha_0, \sigma^2/n)$$

Also

$$\text{Cov}(A_0, A_1) = \text{Cov}(A_0, A_2) = 0$$

$$\text{Cov}(A_1, A_2) = \frac{-A_{12}}{A_{11} A_{22} - A_{12}^2} \sigma^2$$

Then for a predicted value $\hat{Y}(X_1, X_2)$ at any point (x_1, x_2) we have

$$\sigma_{\hat{Y}}^2 = \left[\frac{1}{n} + \left(\frac{A_{22}}{A_{11} B_{22}} \right) x_1^2 - 2 \left(\frac{A_{12}}{A_{11} B_{22}} \right) x_1 x_2 + \left(\frac{A_{11}}{A_{11} B_{22}} \right) x_2^2 \right] \sigma^2$$

This together with

$$\mu_{\hat{Y}} = E(\hat{Y}) = \tilde{Y}(x_1, x_2)$$

and the fact that

$$\hat{Y}(X_1, X_2) \sim N(\mu_{\hat{Y}}, \sigma_{\hat{Y}}^2)$$

can be used to establish confidence intervals for $\tilde{Y}(x_1, x_2)$

CAUTION: In deriving these formulas it was assumed that $\bar{x}_1 = \bar{x}_2 = 0$. For data in which this shift has not been made, the formulae must be adjusted.

Quadratic Model

The quadratic mode

$$\hat{Y} = \alpha_0 + \alpha_1 x + \alpha_2 x^2$$

can be transformed into a linear model with two predictors by the transformation

$$x_1 = x, \quad x_2 = x^2$$

A-3. LINEAR LEAST SQUARES WITH THREE PREDICTORS

Sample:

$$(x_{1i}, x_{2i}, x_{3i}, y_i) \quad i=1, 2, \dots, n$$

Assumptions:

$$A1 \quad \tilde{y}(\tilde{x}_1, \tilde{x}_2, \tilde{x}_3) = \alpha_0 + \alpha_1 \tilde{x}_1 + \alpha_2 \tilde{x}_2 + \alpha_3 \tilde{x}_3$$

$$A2 \quad x_i = \tilde{x}_i \quad i=1, 2, 3$$

$$A3 \quad y_i = \tilde{y}(x_1, x_2, x_3) + e_i$$

Computational Format

n	$\sum x_{1i}$	$\sum x_{2i}$	$\sum x_{3i}$	$\sum y_i$	$A_{uv} = [n(\sum x_{ui} x_{vi}) - (\sum x_{ui})(\sum x_{vi})]$
	$\sum x_{1i}^2$	$\sum x_{1i} x_{2i}$	$\sum x_{1i} x_{3i}$	$\sum x_{1i} y_i$	$u, v=1, 2, 3$
		$\sum x_{2i}^2$	$\sum x_{2i} x_{3i}$	$\sum x_{2i} y_i$	$A_{uy} = [n(\sum x_{ui} y_i) - (\sum x_{ui})(\sum y_i)]$
			$\sum x_{3i}^2$	$\sum x_{3i} y_i$	$B_{uv} = (A_{11} A_{uv} - A_{1u} A_{1v}) / A_{11}$
				$\sum y_i^2$	
	A_{11}	A_{12}	A_{13}	A_{1y}	
		A_{22}	A_{23}	A_{2y}	$C_{uv} = (B_{22} B_{uv} - B_{2u} B_{2v}) / B_{22}$
			A_{33}	A_{3y}	
				A_{yy}	$D_e = (C_{33} C_{yy} - C_{3y}^2) / C_{33}$
		B_{22}	B_{23}	B_{2y}	
			B_{33}	B_{3y}	$S_e = D_e / (n-4) = \frac{1}{n-4} \sum e_i^2$
				B_{yy}	
			C_{33}	C_{3y}	$A_3 = C_{3y} / C_{33}$
				C_{yy}	
				D_e	$A_2 = (B_{2y} - a_3 B_{2y}) / B_{22}$
a_0	a_1	a_2	a_3	S_e^2	

$$a_1 = (A_1 y - a_3 A_{13} - a_2 A_{12}) / A_{11}$$

$$a_0 = (\sum y_i - a_3 \sum x_{3i} - a_2 \sum x_{2i} - a_1 \sum x_{1i}) / n$$

$$\hat{Y}(x_1, x_2, x_3) = a_0 + a_1 x_1 + a_2 x_2 + a_3 x_3$$

Statistics

$$\hat{Y}(x_1, x_2, x_3) = A_0 + A_1 x_1 + A_2 x_2 + A_3 x_3$$

=Prediction Equation

It can be seen that the A_i 's, and hence $\hat{Y}(x_1, x_2, x_3)$ are normally distributed. Determining their means and variances is quite mathematically involved and will be delayed until the vector solution is considered.

Cubic Polynomial

$$\tilde{y}(x) = \alpha_0 + \alpha_1 x + \alpha_2 x^2 + \alpha_3 x^3$$

Transformation

$$x_1 = x, \quad x_2 = x^2, \quad x_3 = x^3$$

$$\tilde{y}(x) = \alpha_0 + \alpha_1 x_1 + \alpha_2 x_2 + \alpha_3 x_3$$

LINEAR LEAST SQUARES WITH k PREDICTORSSample Data:

$$(x_{1i}, \dots, x_{ki}, y_i) \quad i=1, \dots, n$$

Linear Model:

$$\tilde{Y} = \alpha_0 + \alpha_1 x_1 + \dots + \alpha_k x_k$$

Prediction:

$$\hat{Y} = a_0 + a_1 x_1 + \dots + a_k x_k$$

Computational Format

n	Σx_{1i}	Σx_{2i}	\dots	Σx_{ki}	Σy_i
	Σx_{1i}^2	$\Sigma x_{1i} x_{2i}$	\dots	$\Sigma x_{1i} x_{ki}$	$\Sigma x_{1i} y_i$
		Σx_{2i}^2	\dots	$\Sigma x_{2i} x_{ki}$	$\Sigma x_{2i} y_i$
					\vdots
				Σx_{ki}^2	$\Sigma x_{ki} y_i$
					Σy_i^2
<hr/>					
	A_{11}	A_{12}	\dots	A_{1k}	A_{1y}
		A_{22}	\dots	A_{2k}	A_{2y}
				\vdots	\vdots
				A_{kk}	A_{ky}
					A_{yy}
<hr/>					
		$A_{2.22}$	\dots	$A_{2.2k}$	$A_{2.2y}$
		$A_{2.33}$	\dots	$A_{2.3k}$	$A_{2.3y}$
				\vdots	\vdots
				$A_{2.kk}$	$A_{2.yy}$
<hr/>					

$$A_{3 \cdot 3 3} \cdot \dots \cdot A_{3 \cdot 3 k} \quad A_{3 \cdot 3 y}$$

⋮

$$A_{3 \cdot k y}$$

$$A_{k \cdot k k}$$

$$A_{k \cdot k y}$$

$$A_{k \cdot y y}$$

$$D$$

A_0	A_1	A_{k-1}	A_k	$S_e^2 = D_e / (n-k-1)$
-------	-------	-----------	-------	-------------------------

This will be presented for k predictors (x_1, \dots, x_k). Let the sample data be $(x_{1i}, x_{2i}, \dots, x_{ki}, y_i)$ with $i=1, \dots, n$. This data can be presented as a vector \vec{y} and a matrix \vec{x} where

$$\vec{y} = \begin{Bmatrix} y_1 \\ \vdots \\ y_n \end{Bmatrix} \text{ and } \vec{x} = \begin{Bmatrix} 1 & x_{11} & \dots & x_{k1} \\ 1 & x_{12} & \dots & x_{k2} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{1n} & \dots & x_{kn} \end{Bmatrix}$$

Now define vectors $\vec{\alpha}$, \vec{x} and \vec{a} as

$$\vec{\alpha} = \begin{Bmatrix} \alpha_0 \\ \vdots \\ \alpha_k \end{Bmatrix}, \vec{x} = \begin{Bmatrix} x_0 \\ x_1 \\ \vdots \\ x_k \end{Bmatrix}, \text{ and } \vec{a} = \begin{Bmatrix} a_0 \\ a_1 \\ \vdots \\ a_k \end{Bmatrix}$$

where $x_0=1$. The linear model then takes the form

$$\hat{y} = \hat{y}(x_1, \dots, x_k) = \vec{\alpha}^T \vec{x} = \sum_{j=0}^k \alpha_j x_j$$

where $\vec{\alpha}^T$ denotes the row vector which is the transpose of $\vec{\alpha}$, i.e.,

$$\vec{\alpha}^T = (\alpha_0, \alpha_1, \dots, \alpha_k)$$

The fitted equation are

$$\hat{y} = \hat{y}(x_1, \dots, x_k) = \vec{a}^T \vec{x} = \sum_{j=0}^k a_j x_j$$

where the a_j 's are established to minimize

$$D = \sum e_i^2$$

with

$$e_i = y_i - \hat{y}(x_{1i}, \dots, x_{ki}) = y_i - \sum_{j=0}^k a_j x_{ji}$$

In vector form, we have

$$\vec{e} = \vec{y} - \vec{x}\vec{a}$$

so that

$$D = \vec{e}^T \vec{e}$$

The normal equations (to minimize D) are

$$\vec{x}^T \vec{x} \vec{a} = \vec{x}^T \vec{y}$$

with the solutions

$$\vec{a} = (\vec{x}^T \vec{x})^{-1} \vec{x}^T \vec{y}$$

Expressing the coefficients in terms of random variables, we have

$$\vec{A} = (\vec{x}^T \vec{x})^{-1} \vec{x}^T \vec{Y}$$

where

$$\vec{Y}^T = (Y_1, \dots, Y_n) = (\vec{\tilde{Y}} + \vec{E})^T = \vec{\tilde{Y}}^T + \vec{E}^T$$

with

$$\begin{aligned} \vec{E}^T &= (E_1, \dots, E_n) \\ \vec{\tilde{Y}} &= (\tilde{Y}_1, \dots, \tilde{Y}_n) = (\vec{x} \vec{a})^T \\ \vec{\tilde{Y}} &= \vec{\tilde{Y}} + \vec{E} \end{aligned}$$

using

$$\begin{aligned} \mathcal{E}(\vec{E}) &= \vec{0} \\ \mathcal{E}(\vec{E} \vec{E}^T) &= I \sigma^2 \end{aligned}$$

where I is the nxn identity matrix, we have

$$\begin{aligned} \mathcal{E}(\vec{Y}) &= \vec{\tilde{Y}} \\ \mathcal{E}(\vec{Y} \vec{Y}^T) &= I \sigma^2 + \vec{\tilde{Y}} \vec{\tilde{Y}}^T \end{aligned}$$

and hence the covariance matrix for \vec{Y} is

$$\text{Cov}(\vec{Y}, \vec{Y}^T) = \mathcal{E}(\vec{Y} \vec{Y}^T) - \vec{\tilde{Y}} \vec{\tilde{Y}}^T = I \sigma^2$$

Then

$$\begin{aligned} E(\vec{A}) &= (\vec{x}^T \vec{x})^{-1} \vec{x}^T \mathcal{E}(\vec{Y}) \\ &= (\vec{x}^T \vec{x})^{-1} \vec{x}^T \vec{x} \vec{\alpha} = \vec{\alpha} \end{aligned}$$

Thus \vec{a} provides unbiased estimates for the elements of $\vec{\alpha}$.

For the variances and covariances of the coefficients we have

$$\text{Cov}(\vec{A}, \vec{A}^T) = (\vec{x}^T \vec{x})^{-1} \sigma^2$$

Finally, for \hat{Y} at any \vec{x} we have

$$\mathcal{E}(\hat{Y}) = \tilde{y}$$

and

$$\begin{aligned} \sigma_{\hat{Y}}^2 &= \vec{x}^T \text{Cov}(\vec{A}, \vec{A}^T) \vec{x} \\ &= \vec{x}^T (\vec{x}^T \vec{x})^{-1} \vec{x} \sigma^2 \end{aligned}$$

A P E N D I X B

S A M P L E L E A S T S Q U A R E S C A L C U L A T I O N

B-1. S T R A I G H T L I N E R E G R E S S I O N F O R S A M P L E I I

i	t_i'	x_i'	t_i	x_i	\hat{x}_i	e_{xi}
1	872	2183.2	- 5	- 371.2	- 405.9	+ 34.7
2	873	2241.8	- 4	- 312.6	- 324.7	+ 12.1
3	874	2305.5	- 3	- 248.9	- 243.5	- 5.4
4	875	2377.1	- 2	- 177.3	- 162.3	- 15.0
5	876	2451.6	- 1	- 102.8	- 81.2	- 21.6
6	877	2533.8	0	- 20.6	0.0	- 20.6
7	878	2619.6	1	65.2	81.2	- 16.0
8	879	2707.7	2	153.3	162.4	- 9.1
9	880	2799.6	3	245.2	243.6	+ 1.6
10	881	2891.6	4	337.2	324.8	+ 12.4
11	882	2987.3	5	432.9	406.0	+ 26.9
SUM			0	0.4	0.4	0.0
		28098.8				

$$\bar{x} = \frac{1}{n} \sum x_i = 2554.44$$

$$x_i = x_i - \bar{x}$$

$$n=11 \quad \begin{aligned} \sum t_i &= 0 \\ \sum t_i^2 &= 110 \end{aligned}$$

$$\begin{aligned} \sum x_i &= 0.4 \\ \sum t_i x_i &= 8,931.2 \\ \sum x_i^2 &= 728,868.12 \end{aligned}$$

$$A_{11} = 110$$

$$\begin{aligned} A_{1x} &= 8,931.2 \\ A_{xx} &= 728,868.11 \end{aligned}$$

$$A_{ee}=3719.62$$

$$A_0=0.04$$

$$A_1=81.19$$

$$S_e^2 = A_{ee}/(n-2) = 413.29 \quad S_e = 20.33$$

$$\hat{X}(t) = a_0 + a_1 t = 0.04 + 81.19t$$

B-2. QUADRATIC REGRESSION FOR SAMPLE II

$t_{1i}=t_i$	$t_{2i}=t_i^2$	x_i	\hat{x}_i	e_{xi}
- 5	25	- 371.2	-375.1	+ 3.9
- 4	16	- 312.6	-312.4	- 0.2
- 3	9	- 248.9	- 245.6	- 3.3
- 2	4	- 177.3	-174.7	- 2.6
- 1	1	- 102.8	- 99.7	- 3.1
0	0	- 20.6	- 20.5	- 0.1
1	1	65.2	+ 62.7	+ 2.5
2	4	153.3	+150.1	+ 3.2
3	9	245.2	+241.6	+ 3.6
4	16	337.2	+337.1	+ 0.1
5	25	432.9	+436.8	- 3.9
SUM 0	110	0.4	0.3	0.1

$$n=11$$

$$\sum t_{1i}=0$$

$$\sum t_{2i}=110$$

$$\sum X_i=0.4$$

$$\sum t_{1i}^2=110$$

$$\sum t_{1i} t_{2i}=0$$

$$\sum t_{1i} X_i=8,931.2$$

$$\sum t_{2i}^2=1958$$

$$\sum t_{2i} X_i=1769.2$$

$$\sum X_i^2=728,868.1$$

$$A_{11} = 110$$

$$A_{12} = 0$$

$$A_{1x} = 8,931.2$$

$$A_{22} = 858$$

$$A_{2x} = 1765.2$$

$$A_{xx} = 728,868.11$$

$$A_{2,22} = 858$$

$$A_{2,2x} = 1765.2$$

$$A_{2,xx} = 3719.62$$

$$A_{ee} = 87.999$$

$$a_0 = -20.53$$

$$a_1 = 81.19$$

$$a_2 = 2.057$$

$$S_e^2 = A_{ee}/8 = 11.00$$

$$\hat{X}(t) = -20.53 + 81.19t + 2.057t^2$$

$$S_e = 3.32$$

In Appendix A, it was shown that the confidence intervals for $x(t)$ at any time t had the form

$$(\hat{X}(t) - C_j(t) S_e, \hat{X}(t) + C_j(t) S_e) \quad j=1,2$$

where

$$C_1(t) = k_1 \sqrt{\frac{1}{n} + \frac{t^2}{\sum t_i^2}}$$

$$C_2(t) = k_2 \sqrt{\frac{1}{n} + \frac{A_{22}}{A_{11}A_{2.22}} t_1^2 - 2 \left(\frac{A_{12}}{A_{11}A_{2.22}} \right) t_1 (t_2 - \bar{t}_2) + \left(\frac{A_{11}}{A_{11}A_{2.22}} \right) (t_2^2 - \bar{t}_2^2)}$$

are the appropriate terms for the linear and quadratic regression curves, respectively. For a 95% confidence level and $n-2$ or $n-3$ degrees of freedom for the Student-T distribution we find $k_1=1.833$ and $k_2=1.860$. Introducing the numerical values determined in the preceding sections of this appendix, we find

$$C_1(t) = 1.833 \sqrt{\frac{1}{11} + \frac{t^2}{110}}$$

$$C_2(t) = 1.860 \sqrt{\frac{1}{11} + \frac{t^2}{110} + \frac{(t^2 - 10)^2}{858}}$$

The relationships of $C_1(t)$ and $C_2(t)$ and the increments $S_{1e}C_1(t)$ and $S_{2e}C_2(t)$ are shown below using $S_{1e}=20.33$ and $S_{2e}=3.32$.

t	$C_1(t)$	$C_2(t)$	$S_{1e} C_1(t)$	$S_{2e} C_2(t)$
0	.553	.847	11.24	2.81
± 1	.580	.820	11.79	2.72
± 2	.654	.765	13.30	2.54
± 3	.762	.776	15.49	2.58
± 4	.891	.981	18.11	3.26
± 5	1.034	1.417	21.02	4.70

The confidence interval for $x(t)$ is shortest at $t=0$ (the sample midpoint) for the linear regression. To find the value of t in the quadratic regression for which the confidence interval is shortest, consider

$$z = \frac{1}{11} + \frac{t^2}{110} + \frac{(t^2-10)^2}{858}$$

now

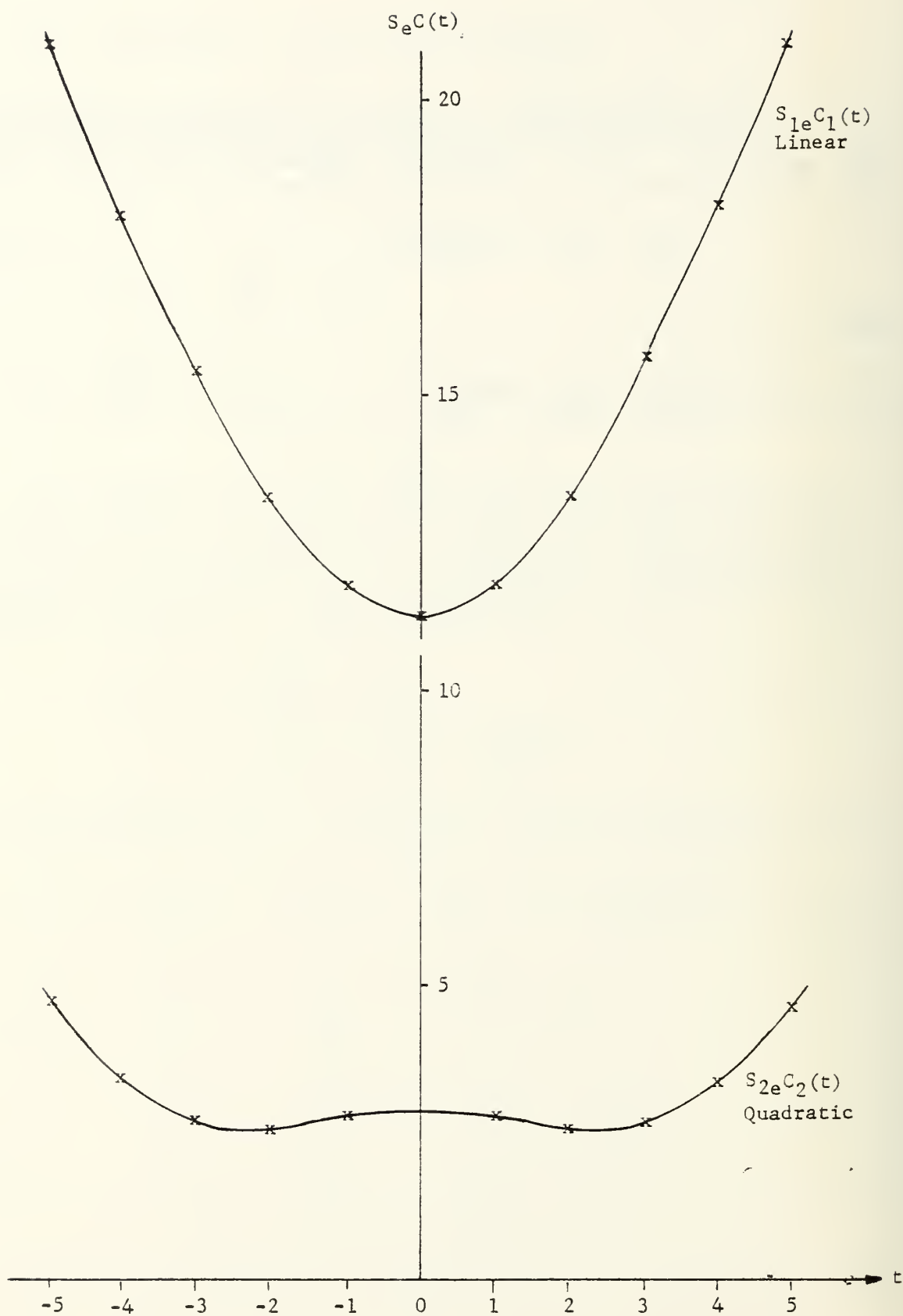
$$\frac{dz}{dt^2} = \frac{1}{110} + \frac{2(t^2-10)}{858} = 0$$

$$220 \quad t^2 = 2200 - 858 = 1342$$

$$t^2 = 6.10$$

$$t = 2.47$$

$$C_2(2.47) = 0.7535$$



DISTRIBUTION LIST

No. of Copies

Commanding Officer Attn: Mr. R. L. Marimon, Code 70 Naval Undersea Warfare Engineering Station Keyport, WA 98345	6
Library, Code 0142 Naval Postgraduate School Monterey, CA 93940	2
Dean of Research Code 012A Naval Postgraduate School Monterey, CA. 93940	1
Prof. J. B. Tysver Code 55Ty Naval Postgraduate School Monterey, CA 93940	10
Naval Undersea Warfare Engineering Station Keyport, WA 98345	
Attn: Code 50	1
Code 51	1
Code 52	1
Code 53	1
Code 54	1
Code 5122	2
Code 0116, Technical File Branch	3
Naval Postgraduate School Monterey, CA. 93940	
Attn: Prof. D.B. Wilson, Code 61W1	3
Prof. H. A. Titus, Code 62Ts	1
Prof. A. R. Washburn, Code 55Ws	1
R. J. Stampfel, Code 55	1

U187371

DUDLEY KNOX LIBRARY - RESEARCH REPORTS



5 6853 01062199 8

010757
NPS